Système intelligent pour le traitement et l'analyse d'anomalies gastro-intestinales dans des séquences de capsules vidéo-endoscopiques

Tan-Sy Nguyen

UNIVERSITÉ
SORBONNE PARIS NORD

École Doctorale Galilée

THÈSE

Pour obtenir le titre de Docteur de l'Université Sorbonne Paris Nord

Spécialité doctorale "Mathématiques"

Soutenue publiquement à l'Université Sorbonne Paris Nord par

Tan Sy NGUYEN

le 18 Décembre 2023

A Smart System for Processing and Analyzing Gastrointestinal Abnormalities in Wireless Capsule Endoscopy

Devant un jury composé de :

M. Faouzi ALAYA CHEIKH,	Norwegian University of Science and Technology	Examinateur
M. Azeddine BEGHDADI,	Université Sorbonne Paris Nord	Examinateur
M. John CHAUSSARD,	Université Sorbonne Paris Nord	Co-encadrant
M. Frédéric DUFAUX,	Université Paris-Saclay	Rapporteur
M. Thuong LE-TIEN,	Ho Chi Minh City University of Technology	Membre invité
Mme Marie LUONG,	Université Sorbonne Paris Nord	Co-encadrant
M. Hatem ZAAG,	Université Sorbonne Paris Nord	Directeur de thèse
M. Habib ZAIDI,	Université de Genève	Rapporteur
Mme Lu ZHANG,	INSA de Rennes	Rapporteur

Laboratoire Analyse, Géométrie et Application

CNRS UMR 7539, Villetaneuse, France

Laboratoire de Traitement et de Transport de l'Information

UR 3043, Villetaneuse, France

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my doctoral supervisors: Prof. Hatem ZAAG, Assoc. Prof. John CHAUSSARD, and Assoc. Prof. Marie LUONG for their guidance and support during my thesis. I am thankful to them for patientially listening to all my questions, sharing their immense knowledge with me and giving me a lot of outstanding advice to overcome scientific obstacles and improve my research competence.

Moreover, I am deeply thankful to Prof. Habib ZAIDI, Prof. Frédéric DUFAUX, and Assoc. Prof. Lu ZHANG, the reviewers of my dissertation. It is a great honor for me to have my Ph.D. thesis evaluated by these experts. I also want to express my gratitude to the other members of my thesis committee. Prof. Azeddine BEGHDADI's insightful comments, valuable suggestions, and interesting questions have significantly improved the quality of my thesis. I am also grateful to Prof. Faouzi ALAYA CHEIKH for his presence as a member of the defense committee.

Next, I extend my sincere thanks to Prof. Thuong LE-TIEN for his encouragement and valuable advice, not only in the scientific activities, but also in my career direction.

I owe my warmest affection to all members of the laboratories LAGA and L2TI of the Université Sorbonne Paris Nord for their kindness and helpfulness, especially the secretaries Yolande (LAGA) and Isabelle (L2TI).

Moreover, I would like to thank all my Vietnamese friends for their encouragement and support. Together, we have shared many wonderful and unforgettable memories.

Finally, I also would like to express my appreciation to my family for their sacrifice and consideration. They are always supporting and encouraging me with their best effort.

Villetaneuse, December 18^{th}, 2023

Tan Sy NGUYEN

Résumé

Dans cette thèse, nous abordons les défis liés à l'identification et au diagnostic des lésions pathologiques dans le tractus gatro-intestinal (GI). L'analyse des quantités massives d'informations visuelles obtenues par une capsule vidéo-endoscopique (CVE) qui est un excellent outil pour visualiser et examiner le tractus GI y compris l'intestin grêle, représente une charge considérable pour les cliniciens, entraînant un risque accru de diagnostic erroné.

Afin de palier à ce problème, nous développons un système intelligent capable de détecter et d'identifier automatiquement diverses pathologies gastro-intestinales. Cependant, la qualité limitée des images acquises en raison de distorsions telles que le bruit, le flou et l'éclairement non uniforme constitue un obstacle significatif. Par conséquent, les techniques de prétraitement des images jouent un rôle crucial dans l'amélioration de la qualité des images acquises, facilitant ainsi les tâches de haut niveau telles que la détection et la classification des anomalies. Afin de résoudre les problèmes liés à la qualité limitée des images causée par les distorsions mentionnées précédemment, plusieurs nouveaux algorithmes d'apprentissage ont été proposés. Plus précisément, les avancées récentes dans le domaine de la restauration et de l'amélioration de la qualité des images reposent sur des approches d'apprentissage qui nécessitent des paires d'images déformées et de référence pour l'entraînement. Cependant, en ce qui concerne la CVE, un défi significatif se pose en raison de l'absence d'une base de données dédiée pour évaluer la qualité des images. À notre connaissance, il n'existe actuellement aucune base de données spécialisée conçu spécifiquement pour évaluer la qualité vidéo en CVE. Par conséquent, en réponse à la nécessité d'une base de données complète d'évaluation de la qualité vidéo, nous proposons tout d'abord la "Base de données axée sur la qualité pour l'endoscopie vidéo par capsule/ Quality-Oriented Database for Video Capsule Endoscopy" (QVCED).

Ensuite, nos résultats montrent que l'évaluation de la gravité des distorsions améliore significativement l'efficacité de l'amélioration de l'image, en particulier en cas d'illumination inégale. À cette fin, nous proposons une nouvelle métrique dédiée à l'évaluation et à la quantification de l'éclairage inégal dans les images laparoscopiques ou par CVE, en extrayant l'éclairement de l'arrière-plan de l'image et en tenant compte de l'effet de la mise en égalisation de l'histogramme. Notre métrique offre une performance supérieure à celle de certaines méthodes d'évaluation les plus avancées de la

qualité d'image sans référence (NR-IQA), démontrant sa supériorité et sa performance compétitive par rapport aux méthodes d'évaluation de la qualité d'image avec référence complète (FR-IQA).

Après avoir effectué l'étape d'évaluation, nous développons une méthode d'amélioration de la qualité d'image visant à améliorer la qualité globale des images. Le nouvel algorithme est basé sur un mécanisme de l'attention croisée, qui permet d'établir l'interaction d'information entre la tâche de l'extraction du niveau de distorsion et de la localisation de régions dégradées. En employant cet algorithme, nous sommes en mesure d'identifier et de cibler précisément les zones spécifiques des images affectées par les distorsions. Ainsi, cet algorithme permet le traitement approprié adapté à chaque région dégradée, améliorant ainsi efficacement la qualité de l'image.

Suite à l'amélioration de la qualité de l'image, des caractéristiques visuelles sont extraites et alimentées dans un classificateur pour fournir un diagnostic par classification. La difficulté dans le domaine de CVE est qu'une partie significative des données reste non étiquetée. Pour relever ce défi, nous avons proposé une méthode efficace basée sur l'approche d'apprentissage auto-supervisé ("Self-Supervised Learning" ou SSL en anglais) afin d'améliorer les performances de la classification. La méthode proposée, utilisant le SSL basé sur l'attention, ont réussi à résoudre le problème des données étiquetées limitées couramment rencontré dans la littérature existante.

Mots-Clés — Capsule Vidéo-Endoscopique (CVE), évaluation de la qualité d'image sans référence, égalisation d'histogramme, base de données de distorsions, amélioration de la qualité de l'image, algorithme de l'attention croisée, classification basée sur l'apprentissage auto-supervisé (SSL).

Abstract

In this thesis, we address the challenges associated with the identification and diagnosis of pathological lesions in the gastrointestinal (GI) tract. Analyzing massive amounts of visual information obtained by Wireless Capsule Endoscopy (WCE) which is an excellent tool for visualizing and examining the GI tract (including the small intestine), poses a significant burden on clinicians, leading to an increased risk of misdiagnosis.

In order to alleviate this issue, we develop an intelligent system capable of automatically detecting and identifying various GI disorders. However, the limited quality of acquired images due to distortions such as noise, blur, and uneven illumination poses a significant obstacle. Consequently, image pre-processing techniques play a crucial role in improving the quality of captured frames, thereby facilitating subsequent high-level tasks like abnormality detection and classification. In order to tackle the issues associated with limitations in image quality caused by the aforementioned distortions, novel learning-based algorithms have been proposed. More precisely, recent advancements in the realm of image restoration and enhancement techniques rely on learning-based approaches that necessitate pairs of distorted and reference images for training. However, a significant challenge arises in WCE which is an excellent tool for visualizing and diagnosing GI disorders, due to the absence of a dedicated dataset for evaluating image quality. To the best of our knowledge, there currently exists no specialized dataset designed explicitly for evaluating video quality in WCE. Therefore, in response to the need for an extensive video quality assessment dataset, we first introduce the "Quality-Oriented Database for Video Capsule Endoscopy" (QVCED).

Subsequently, our findings show that assessing distortion severity significantly improves image enhancement effectiveness, especially in the case of uneven illumination. To this end, we propose a novel metric dedicated to the evaluation and quantification of uneven illumination in laparoscopic or WCE images, by extracting the image's background illuminance and considering the mapping effect of Histogram Equalization. Our metric outperforms some state-of-the-art No-Reference Image Quality Assessment (NR-IQA) methods, demonstrating its superiority and competitive performance compared to Full-Reference IQA (FR-IQA) methods.

After conducting the assessment step, we proceed to develop an image quality enhancement method aimed at improving the overall quality of the images. This is

achieved by leveraging the cross-attention algorithm, which establishes a comprehensive connection between the extracted distortion level and the degraded regions within the images. By employing this algorithm, we are able to precisely identify and target the specific areas in the images that have been affected by distortions. This allows an appropriate enhancement tailored to each degraded region, thereby effectively improving the image quality.

Following the improvement of image quality, visual features are extracted and fed into a classifier to provide a diagnosis through classification. The challenge in the WCE domain is that a significant portion of the data remains unlabeled. To overcome this challenge, we have developed an efficient method based on the self-supervised learning (SSL) approach to enhance the performance of classification. The proposed method, utilizing attention-based SSL, has successfully addressed the issue of limited labeled data commonly encountered in the existing literature.

Keywords — Wireless Capsule Endoscopy (WCE), No-Reference image quality assessment (NR-IQA), Histogram Equalization, distortion dataset, image quality enhancement, cross-attention algorithm, Self-Supervised Learning (SSL)-based classification.

Contents

Acknowledgments iii

Résumé v

Abstract vii

Contents viii

List of Tables xvi

List of Figures xviii

List of Notations xxiv

Chapter 1: Introduction 1

1 Wireless Capsule Endoscopy (WCE) imaging modality 2

2 Objectives 5

3 Contributions 7

3.1 WCE dataset creation (QVCED) 8

3.2 Image quality assessment metric (IHED) 9

3.3 Image quality enhancement method (TCFA) 9

3.4 Classification method (DWSA-SSL) 10

4 Thesis organization 10

5 Publications 11

Chapter 2: Related works **13**

1 Introduction . . . 15

2 Image definition . . . 15

3 Distortion models . . . 16

3.1 Additive Gaussian noise model . . . 16

3.2 Blur model . . . 17

3.2.1 Motion blur model . . . 17

3.2.2 Defocus blur model . . . 18

3.3 Uneven illumination model . . . 18

4 The State-of-the-art (SOTA) approaches . . . 20

4.1 Introduction . . . 20

4.2 The SOTA Image Quality Assessment (IQA) metrics . . . 21

4.2.1 Introduction . . . 21

4.2.2 Full-Reference IQA (FR-IQA) . . . 21

4.2.2.1 Mathematical metrics . . . 21

4.2.2.2 Human Visual System-based metrics . . . 22

4.2.2.3 Other metrics . . . 24

4.2.3 Reduced-Reference IQA (RR-IQA) . . . 25

4.2.4 No-Reference IQA (NR-IQA) . . . 26

4.2.4.1 Distortion-specific NR-IQA methods . . . 27

4.2.4.2 General-purpose NR-IQA methods . . . 29

4.2.5 metrics . . . 34

4.2.5.1 PLCC(Pearson Linear Correlation Coefficient) 34

4.2.5.2 SROCC(Spearman Rank-Ordered Correlation Coefficient) . . . 34

4.2.5.3 KROCC(Kendall Rank Order Correlation Coeffcient) . . . 34

4.2.5.4 RMSE(Root Mean Squared Error) 35
4.2.6 Discussion 35
4.3 The SOTA image quality enhancement methods 35
4.3.1 Introduction 35
4.3.2 The SOTA image denoising methods 36
4.3.2.1 Classical methods 36
4.3.2.2 Dictionary learning-based methods 39
4.3.2.3 Deep-learning-based methods 40
4.3.3 The SOTA image deblurring methods 42
4.3.3.1 Classical methods 42
4.3.3.2 Deep-learning-based methods 43
4.3.4 The SOTA uneven illumination correction methods . . 44
4.3.4.1 Classical methods 44
4.3.4.2 Deep-learning-based methods 46
4.3.5 Discussion 47
4.4 The SOTA WCE image classification methods 48
4.4.1 Introduction 48
4.4.2 Mono-pathology classification 48
4.4.3 Multi-pathology classification 50
4.4.4 Discussion 52
5 Conclusion 52

Chapter 3: The proposed distortion dataset: *A Quality-Oriented Database for Video Capsule Endoscopy (QVCED)* **54**

1 Introduction 56
2 Proposed Dataset - QVCED 57
2.1 Reference Videos Selection 58

2.1.1 Noise Assessment 59
2.1.2 Blur Assessment 59
2.1.3 Uneven Illumination Assessment 60
2.2 Distortion Generation 60
2.2.1 Noise Model 60
2.2.2 Defocus Blur Model 62
2.2.3 Motion Blur Model 62
2.2.4 Uneven Illumination Model 64
3 Dataset Analysis 67
3.1 Subjective Test 69
3.1.1 Testing Environment 69
3.1.2 Video Quality Score 70
3.2 Diversity Data Analysis 72
4 Conclusion 73

Chapter 4: The proposed IQA metric: *A No-Reference Measure for Uneven Illumination Assessment on Laparoscopy and WCE images* **74**

1 Introduction 76
2 Proposed Method 82
2.1 Color Space Conversion 82
2.2 Background Illuminance Estimation 83
2.3 Uneven Illumination Assessment 84
3 Performance Evaluation on laparoscopic images 85
3.1 Dataset 86
3.2 Correlation with Human Quality Judgment 86
4 Performance Evaluation on WCE images 87
4.1 Correlation with Human Quality Judgment 87

5 Conclusion 89

Chapter 5: The proposed image quality enhancement method: *TCFA: Triplet Clustering Fusion Autoencoder for Quality Enhancement of WCE images* **90**

1 Introduction 92

2 Related Works 94

2.1 Denoising 95

2.2 Deblurring 95

2.3 Uneven Illumination Correction 96

3 Proposed Method 97

3.1 Overall Pipeline 97

3.2 Distorted Image Projector (DIP) 98

3.3 Distortion Level Encoder (DLE) 99

3.4 Variational Cross-Attention Module (VCAM) 99

3.5 Loss Function 100

4 Experiments 101

4.1 Dataset and Implementation Details 102

4.2 Ablation Study 103

4.2.1 Triplet Loss Margin 103

4.2.2 Effect of each TCFA's component 104

4.3 Comparison With State-of-the-Arts 106

4.3.1 Image Quality Assessment Comparison 106

4.3.2 Visual Quality Comparison 112

4.3.3 Statistical Comparison 112

5 Conclusion 118

Chapter 6: The proposed image classification method: *Dilated Window-based Self-Attention Self-Supervised Learning for Classification in WCE* **120**

1 Introduction . . . 122

2 Related Works . . . 124

3 Basic concept of Self-Supervised Learning . . . 127

4 Proposed Method . . . 131

4.1 Algorithm . . . 132

4.1.1 Training phase . . . 133

4.1.2 Testing phase . . . 134

4.1.3 Loss function . . . 134

4.2 S-DWSA Transformer backbone . . . 136

4.2.1 S-DWSA Transformer block . . . 136

4.2.2 Shifted-Dilated Window-based Self-Attention . . . 138

5 Experimental results . . . 140

5.1 Experimental setting . . . 140

5.1.1 Optimization . . . 140

5.1.2 Dataset . . . 141

5.1.3 Evaluation metrics . . . 143

5.2 Ablation study . . . 144

5.2.1 Ablation on hyper-parameters . . . 145

5.2.2 Ablation on scale-soft-cosine attention . . . 147

5.2.3 Ablation on the Shifted-Dilated Window-based Self-Attention . . . 149

5.3 Comparison with state-of-the-art methods . . . 150

6 Conclusion . . . 152

Chapter 7: Conclusion and perspective **154**

1 Summary and conclusion 154

2 Future work and perspectives 156

Bibliography **158**

List of Tables

3.1 Summary of the proposed wireless capsule endoscopy video quality assessment dataset. 69

4.1 The correlation comparison on the LVQ dataset. 86

4.2 The correlation comparison on the QVCED dataset. 88

5.1 Quantitative comparison with state-of-the-art denoising methods in terms of five image quality assessment metrics 107

5.2 Quantitative comparison with state-of-the-art deblurring methods in terms of five image quality assessment metrics. 108

5.3 Quantitative comparison with state-of-the-art uneven illumination correction methods in terms of six image quality assessment metrics. 109

5.4 Pairwise Comparison technique - conversion scale. 118

6.1 Ablation study on the drop path rates of online and target encoders, while holding other parameters at default including queue size $K = 2048$, temperature $\tau = 0.1$, momentum $\alpha = 0.99$, regular parameter $\beta = 0.75$. 145

6.2 Ablation study on the temperature τ, while holding other parameters at default including queue size $K = 2048$, momentum $\alpha = 0.99$, regular parameter $\beta = 0.75$ and online encoder drop rate of 0.1 146

6.3 Ablation study on the queue size, while holding other parameters at default including temperature $\tau = 0.1$, momentum $\alpha = 0.99$, regular parameter $\beta = 0.75$ and online encoder drop rate of 0.1. 147

6.4 Influence of the scale-soft-cosine attention logits in the proposed architecture on the classification performance. 148

6.5 Comparison of different SSL architectures. 149

6.6 Influence of the window-level attention in the proposed architecture on the classification performance. 150

6.7 Performance comparison with state-of-the-art methods. 150

6.8 Model complexity comparison with state-of-the-art methods. 151

List of Figures

1.1 Capsule endoscopy camera. (source: https://www.mayoclinic.org/tests-procedures/capsule-endoscopy/about/pac-20393366) 3

1.2 A typical Wireless Capsule Endoscopy system. 4

1.3 Sample WCE images with common pathologies: from left to right and top to bottom: Erythema (ERY), Angiectasias (ANG), Blood-Fresh (BF), Blood-Hematin (BH), Erosion (ERO), Ulcers (ULC), Lymphangiectasia (LYM), and Polyp (PYL) [10]. 5

1.4 The proposed system contains two main components of the proposed method: pre-processing and classification. The pre-processing stage involves various proposed techniques to enhance the quality of the input WCE images. The classification component employs advanced algorithms to classify the pre-processed data into distinct pathologies. 6

2.1 Field of View (FOV) measurement of a capsule endoscope: (a) illustration of FOV_{WS} (window surface) and FOV_{EP} (Entrance Pupil), (b) a low-resolution image ($4cm \times 4cm$) on a capsule recorder screen [23]. 19

2.2 (a) Light-shading images when the light source is not directly aligned with the object [24], consider a scenario where the light source is positioned on the surface of a spherical object, with the object in the center of this sphere. (b) The highest intensity point (marked in green) within a sub-image corresponds to the point where the microlens center (marked in blue) is orthogonally projected. The actual camera's central view is marked as the ground truth point (marked in red). This ground truth point can also be alternatively represented using a mathematical model as the reference point (marked in yellow), and the weighted average point (marked in violet) includes a weighted average of pixel intensities within the sub-image. 20

2.3 A standard framework of the Image Quality Assessment (IQA) system based on the Human Visual System (HVS) involves a pre-processing phase preceding the channel decomposition. This pre-processing phase encompasses operations such as alignment, conversion of color spaces, and low-pass filtering through the point-spread function (PSF) to simulate the effects of eye optics. 23

3.1 The frames extracted from reference videos in the QVCED dataset represent a diverse range of findings. 58

3.2 The results of adding AWGN into the reference image with 6 different levels. 61

3.3 The results of adding defocus blur into the reference image with 5 different levels. 63

3.4 Visual representation of the blurring kernel, characterized by a defocus blur standard deviation of $\sigma_{db} = 1$ and motion blur with $L_{mb} = 20, \theta_{mb} = \frac{\pi}{4}$, respectively. 63

3.5 The results of adding motion blur into the reference image with 4 different levels. Fig. (d)-(e) shows us an example of changing the direction from 0 degrees to 45 degrees. 64

3.6 The results of converting the image from RGB to HSV color space. . . . 65

3.7 Gradient masks to simulate the Uneven Illumination which is represented mathematically as a hybrid distribution. 66

3.8 Some examples of the artificial circular-gradient Uneven Illumination mask. 67

3.9 The results of adding uneven illumination into the reference image with 4 different levels and 3 possible positions. 68

3.10 WCE subjective test window. 70

3.11 Age and the processing time distributions of observers participated in the subjective experiment. 71

3.12 Comparison of the subjective score regarding experts and non-experts. . 71

3.13 t-SNE visualization of the embedded feature generated from 20 reference videos by VGG-16 pre-trained network. 72

3.14 Distribution of Variations and Entropy among 20 reference videos. . . . 73

4.1 The AGIC result on the image distorted by 4 levels of uneven illumination. The experiment was operated on LVQ dataset [257]. The column (a) is the original image I. Column (b) is the extracted V-channel V from the original image I. The column (c) is the IB of V which is extracted using LPF ($X = LPF(V)$). The fourth column is the corresponding AGIC value of each IB. 78

4.2 The AGIC results on the synthesis image distorted by 4 levels of uneven illumination. Images (a) to (e) are the uneven illumination mask with a variant gradient level. From (a) to (d) the level of uneven illumination was increased one by one. The image (e) has the same gradient level compared to (d) but instead of using the same position, the mask is created with a different position to have a mask with a larger region of dark which is expected as an inefficient case of the AGIC. The corresponding AGIC indexes are at the top of each image using equation (4.3). 79

4.3 Some image examples used in [93]. Row (i) is the original image. Row (ii) is the corresponding V-channel when they convert the image from RGB color space into HSV color space. 80

4.4 The SDV result on the LVQ [257] dataset. Row (i) is the original image. Row (ii) is the corresponding V-channel when they convert the image from RGB color space into HSV color space. The corresponding SDV values are on the top of each column. 81

4.5 Block diagram of the proposed uneven illumination assessment method. 82

4.6 Original images at different UI levels ((a)-(c)).The corresponding BI maps of **V** component ((d)-(f)). The equalized BI images (**E** signal) ((g)-(i)). The difference signals **D** between the BI maps (BI of the original and the corresponding equalized version) ((j)-(l)). 85

4.7 Scatter plots of MOS versus prediction of IHED on the LVQ dataset. IHED versus the Expert MOS (left). IHED versus Non-Expert MOS (right). 87

4.8 Scatter plots of MOS versus prediction of IHED on the QVCED dataset. IHED versus the Expert MOS (left). IHED versus Non-Expert MOS (right). 88

5.1 (a) Overview of the proposed TCFA. Shaded circles represent processed images, bullet shapes represent deterministic functions, • represents a bifurcation, and red boxes represent cost terms. Dotted lines exclusively indicate the data path which is available only during the training phase. (b) The encoder architecture used in the DIE and DLE. (c) The employed decoder architecture used in the DID. 97

5.2 Quantifying discrimination in clustering results using Davies-Bouldin score and Silhouette score. 104

5.3 Visual comparisons obtained from the ablation study conducted on TCFA. Rows 1 to 3 represent different elements, including the distorted image inputs, the reference image, outputs from TCFA without the VCAM, outputs from TCFA without the Distortion Level Encoder, and outputs from the final version of TCFA, respectively. The visual assessment focuses on addressing issues such as noise, blur, and uneven illumination. Notably, the final version of TCFA demonstrates the most significant improvement in mitigating these aforementioned issues and produces visually appealing results. For a closer examination of the image details, please zoom in. 105

5.4 Quantitative comparison with state-of-the-art denoising methods in terms of the quantity test on BRISQUE/NIQE. 108

5.5 Quantitative comparison with state-of-the-art deblurring methods in terms of the quantity test on BRISQUE/NIQE. 109

5.6 Quantitative comparison with state-of-the-art uneven illumination correction methods in terms of the quantity test on BRISQUE/NIQE/LOE. 110

5.7 Effect of the distortion level on the performances for different enhancement methods. The first column shows a performance comparison among different denoising methods. The second column highlights the performance of deblurring methods, while the third column presents the comparative results for correcting uneven illumination. 111

5.8 Visual comparison of the reconstruction methods of denoising. 113

5.9 Visual comparison of the reconstruction methods of deblurring. 114

5.10 Visual comparison of the reconstruction methods of uneven illumination correction. 115

5.11 Statistical Comparison of different enhancement methods. The first column shows a comparison among different denoising methods. The second column highlights the performance of deblurring methods, while the third column represents the comparative results for correcting uneven illumination. The first row and the second row show the box plots of the results from different enhancement methods while the third row illustrates the pairwise comparison of the corresponding method on different IQA metrics. In each box plot, the number of outliers related to each considered method is also presented. 116

6.1 A model is first trained with a pretext task with unlabeled data, then fine-tuned on the downstream task with a limited amount of labeled data. Usually, convolution layers, which are mostly responsible for learning representations, are transferred. A few fully connected layers towards the end are changed or retrained. 127

6.2 A simple framework for contrastive learning of visual representations. . 128

6.3 The algorithm for SimCLR [300]. 129

6.4 The model architecture of BYOL. After training, we only care about f_θ producing representation, $y = f_\theta(x)$, and everything else is discarded. sg means stop gradient. 130

6.5 Illustration of how Momentum Contrast (MoCo) learns visual representations. 131

6.6 The pipeline of the proposed DWSA-SSL method. $\mathcal{T}_1, \mathcal{T}_2$ are image augmentations. $\mathbf{x}', \mathbf{x}''$ are two augmented views from $\mathbf{x}$ by applying respectively $\mathcal{T}_1, \mathcal{T}_2$. $\mathbf{y}', \mathbf{y}'_w$ and $\mathbf{y}'', \mathbf{y}''_w$ are the online/ target representations corresponding to the patch-level and window-level attention by S-DWSA Transformer backbones, respectively. $\mathbf{z}', \mathbf{z}'_w$ are the online projections of $\mathbf{y}', \mathbf{y}'_w$, respectively. $\mathbf{q}', \mathbf{q}'_w$ are the online prediction of $\mathbf{z}', \mathbf{z}'_w$, respectively. $\mathbf{k}'', \mathbf{k}''_w$ are the target projections (key) of $\mathbf{y}'', \mathbf{y}''_w$, respectively. EMA: Exponential Moving Average. *sg* means stop-gradient. 132

6.7 Overview of SSL approaches. 134

6.8 The architecture of the S-DWSA Transformer. 136

6.9 Two successive S-DWSA Transformer Blocks. 137

6.10 An illustration of the Shifted-Dilated Window-based Self-Attention. In layer l, a regular window partitioning scheme is adopted, and self-attention is computed within each window (patch-level). On the right, the window-level attention is calculated based on 9 dilated surrounding windows. In layer $l+1$, the window partitioning is shifted, and the self-attention computation in the new windows crosses the boundaries of the previous windows in layer l, providing overlapping among them. 139

6.11 The number of images in the various Kvasir-Capsule labeled image classes. 143

6.12 Eight pathological classes in the Kvasir-Capsule dataset [10]. 144

6.13 The Signal Propagation Plot for various attention logits. 148

List of Notations

The notations used in this thesis follow the principles below.

i. The bold capital letters are used to represent a two-dimensional (2D) arrays, which can be a matrix (attention matrix $\mathbf{K}$, etc.), a dictionary of K column vectors (e.g. $\mathbf{D}$) or a 2D image with pixels arranged in rows and columns such as $\mathbf{X}, \mathbf{Y}$.

ii. The bold letter, e.g. $\mathbf{x}$, $\mathbf{y}$, represents a column vector or a vectorization of a 2D image obtained by raster scan the image from top to bottom and left to right.

iii. The subscripts below a symbol stand for the subsets. For instance, $\mathbf{x}_i$ is a sub-column vector (also called image patch) of a vectorization image $\mathbf{x}$, $\mathbf{X}_i$ indicates a 2D array where each of its column is an image patch, etc.

iv. The superscripts express the modes or state of a variable. For example, $\mathbf{x}^{ref}$ is a standard (high quality) image, etc.

v. The array indexing, which encloses the indices in parentheses, is used to access an element of the array. E.g. $\mathbf{X}(i,j)$ is the pixel at i-th row and j-th column of 2D image $\mathbf{X}$, $\mathbf{x}_i(j)$ is the j-th pixel of the column image patch $\mathbf{x}_i$, etc.

The list of important notations is described as follows.

R^n	A n-dimensional space.
$\mathbb{N}$	A set of natural numbers.
$R^{n\times K}$	A space of two-dimensional arrays of n rows and K columns.
$\mathbf{I}$	An image feature represented as a multi-dimentional array.
$\mathbf{X}(i,j)$	A pixel in the feature $\mathbf{X}$ at i-th row and j-th column.
$\mathbf{x}_i$	A column vector or an feature patch represented as a column.
$\mathbf{x}_i(j)$	The j-th element (pixel) of the (column) feature patch $\mathbf{x}_i$.
Ω_α	A vector space in R^K generated by K basis vectors.

$\boldsymbol{\alpha}_i$	A vector of representation coefficients of an image patch $\mathbf{x}_i$ in the vector space Ω_α.
$\mathcal{K}(\cdot)$	A kernel function.
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with mean μ and variance σ^2.
σ	Standard deviation.
μ	Mean.
Σ	Summation.
$\odot$	Element-wise multiplication.
$\|\cdot\|_F$	Frobenius norm.
$\mathbb{E}[\cdot]$	Expectation.
$p(\cdot)$	Probability density function.
k-NN	k-nearest neighbor.
BM3D	Block Matching 3D-based denoising.
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator.
CNN	Convolutional Neural Network.
DE	Decrease of Entropy.
FR	Full-Reference IQA.
IQA	Image Quality Assessment.
MSE	Mean Square Error.
NIQE	Naturalness Image Quality Evaluator.
NR	No-Reference IQA.
PDF	Probability Density Function.
PLCC	Pearson Linear Correlation Coefficient.
PSNR	Peak Signal-to-Noise Ratio.
RR	Reduced-Reference IQA.
SROCC	Spearman Rank Order Correlation Coefficient.
SSIM	Structural Similarity.
VIF	Visual Information Fidelity.
WCE	Wireless Capsule Endoscopy.

CHAPTER 1

Introduction

Chapter content

1 **Wireless Capsule Endoscopy (WCE) imaging modality** **2**

2 **Objectives** **5**

3 **Contributions** **7**

3.1 WCE dataset creation (QVCED) 8

3.2 Image quality assessment metric (IHED) 9

3.3 Image quality enhancement method (TCFA) 9

3.4 Classification method (DWSA-SSL) 10

4 **Thesis organization** **10**

5 **Publications** **11**

This work is devoted to the development of smart computer-aided solutions for the detection and diagnosis of diseases related to the gastrointestinal (GI) tract, which, together with the biliary system, forms the digestive system.

The diagnosis and management of GI abnormalities and disorders pose significant challenges in the field of medicine, particularly in healthcare systems worldwide. GI diseases and disorders affect the GI tract and its various organs, including mouth, esophagus, stomach, small intestine, large intestine, and anus. Several pathologies (or conditions) which can manifest in diverse symptoms, include inflammatory bowel diseases (such as ulcerative colitis, Crohn's disease), gastroesophageal reflux disease, liver cirrhosis, and gastrointestinal cancers. The impact of digestive diseases is staggering, with statistics indicating their widespread prevalence. An alarming increase in the occurrence of several digestive diseases with statistics indicating their widespread prevalence, has had a major impact worldwide. For instance, according to the World Gastroenterology Organization, digestive diseases account for approximately 8 million deaths globally each year, with gastrointestinal cancers alone causing over 3 million deaths annually [1]. In Europe, GI and liver disorders are responsible for about one million deaths each year across Europe with significant healthcare costs since 2000 [2]. Additionally, digestive diseases are responsible for a significant number of hospitalizations and outpatient visits, placing a substantial burden on healthcare resources and budgets. Addressing the burden of GI diseases requires a comprehensive approach that encompasses prevention, early detection, accurate diagnosis, and effective management strategies. Moreover, the complexity of the GI system, coupled with the diverse range of potential abnormalities, necessitates accurate identification and precise diagnosis for effective treatment [3], [4].

1 Wireless Capsule Endoscopy (WCE) imaging modality

Traditional endoscopic procedures have been employed for a long time to visualize and assess the GI tract; however, these techniques are invasive, uncomfortable, and limited in their ability to comprehensively examine the entire digestive system [5]. In recent years, the emergence of Wireless Capsule Endoscopy (WCE) has revolutionized the field of gastroenterology, becoming one of the best tools for visualizing and diagnosing GI disorders (e.g. bleeding, polyps, ulcers) and more interestingly small intestinal diseases. WCE involves the ingestion of a small capsule equipped with a miniaturized camera (Fig. 1.1) that captures images when it traverses the GI tract [6], [7]. This non-invasive and patient-friendly diagnostic modality provides clinicians with an unprecedented view of the entire gastrointestinal system, enabling the identification and evaluation of various

abnormalities, including bleeding, ulcers, tumors, and inflammatory diseases [6].

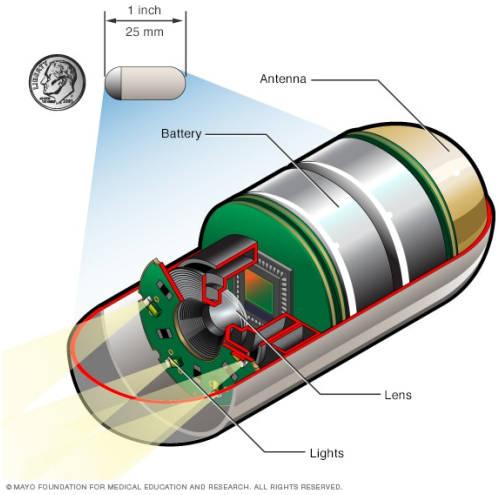


Figure 1.1: *Capsule endoscopy camera. (source: https://www.mayoclinic.org/tests-procedures/capsule-endoscopy/about/pac-20393366)*

By offering detailed visualization and analysis of the mucosal lining, WCE enables the identification of lesions and abnormalities that might otherwise go unnoticed [6]. Furthermore, WCE eliminates the need for sedation or anesthesia typically required for traditional endoscopic procedures, making it more accessible and comfortable for patients [8]. The real-time and continuous monitoring capabilities of WCE facilitate the observation of dynamic changes within the GI system, aiding in the timely diagnosis of time-sensitive conditions and guiding appropriate treatment strategies [7].

In WCE, after several hours of fasting, the patient ingests a vitamin-sized electronic pill, which passes through the GI tract by peristalsis. While traveling through the GI tract, the pill takes images and transmits image data wirelessly to a portable data logger unit attached to a belt, around the patient's waist. During the procedure, patients are free to conduct their daily activities such as walking, sitting, driving, etc. However, the patient should avoid strenuous physical activity, especially if it involves sweating, and should not bend or stoop during the procedure. After 8 - 10 hours, the battery life of the capsule runs out and the image data stored in the data logger are transferred to a workstation or a personal computer (PC) where the images are reconstructed and displayed for medical diagnosis. Generally, the capsule comes out from the body naturally after two to three days [9]. The typical WCE system is shown in Fig. 1.2.

WCE has introduced a new set of challenges in handling the vast amount of visual data generated during the examination of the gastrointestinal tract. First of all, the

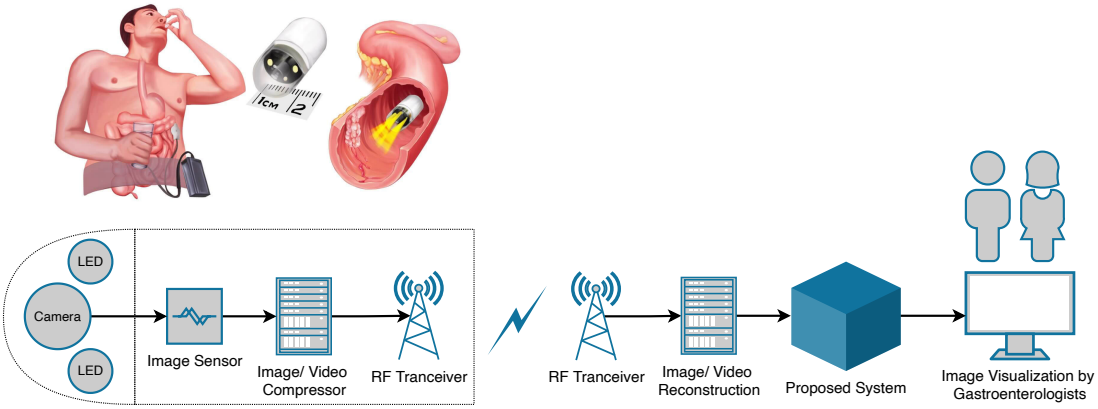
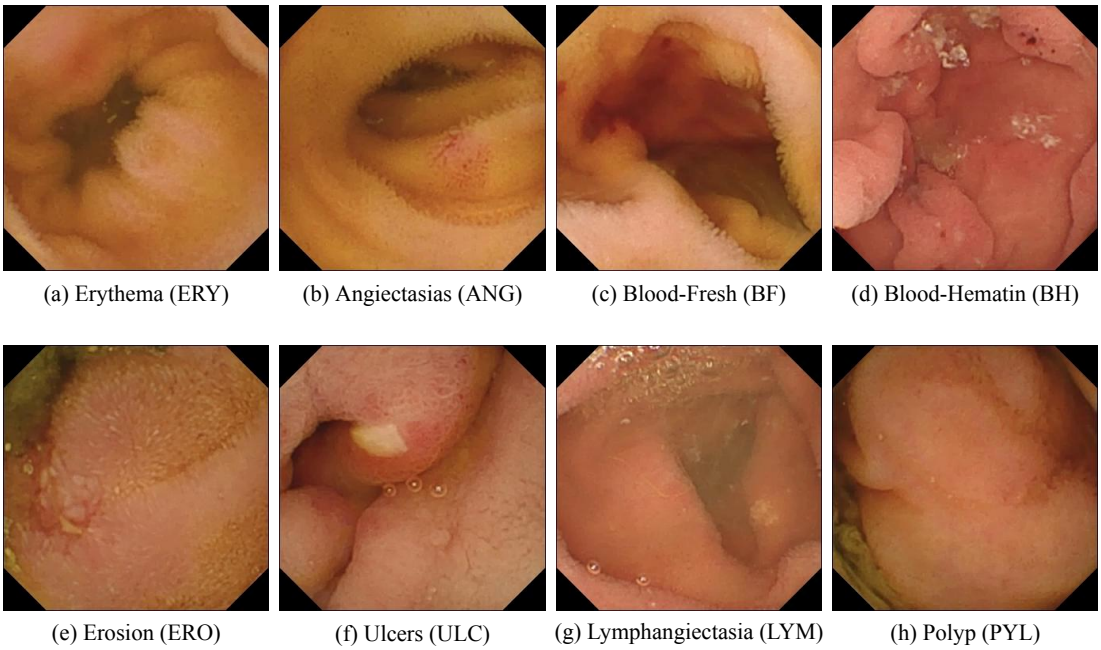


Figure 1.2: *A typical Wireless Capsule Endoscopy system.*

continuous capture of high-resolution images and videos by the WCE device results in a significant volume of data that needs to be effectively managed and analyzed. For example, a single clinical procedure can generate 50,000 images and hours of video footage [11]. This poses challenges in terms of data storage capacity, transmission bandwidth, and computational resources required for processing. The sheer volume of data necessitates efficient data management techniques to store, organize, and retrieve the information effectively. Furthermore, ensuring the quality and reliability of the captured images and videos becomes crucial for accurate diagnosis and interpretation. The major issue with WCE images is their quality. Indeed, the capsule's motion in the GI tract is uncontrolled, and images are taken under low illumination. Some common WCE image artifacts include noise, motion blur, and uneven illumination can greatly impact the overall quality and usability of the visual data, potentially leading to misinterpretation or missed abnormalities. Therefore, it is imperative to develop robust and efficient techniques for data capacity management and quality analysis to extract meaningful information from the extensive WCE datasets.

However, addressing the challenge of effectively extracting meaningful information from these extensive datasets necessitates the development of an advanced smart solution to effectively handle and analyze the massive amount of visual data generated by WCE [7]. These innovations will unlock the full potential of WCE technology in the field of gastroenterology, enabling more accurate diagnoses and improved patient care [12].

The remainder of this chapter is organized as follows. Section 2 presents the problem motivation and context. Subsequently, Section 3, we will summarize the significant contributions made by this thesis. In section 4, we briefly describe the organization of our thesis. Finally, the publications are presented in Section 5.



(a) Erythema (ERY) (b) Angiectasias (ANG) (c) Blood-Fresh (BF) (d) Blood-Hematin (BH)

(e) Erosion (ERO) (f) Ulcers (ULC) (g) Lymphangiectasia (LYM) (h) Polyp (PYL)

Figure 1.3: *Sample WCE images with common pathologies: from left to right and top to bottom: Erythema (ERY), Angiectasias (ANG), Blood-Fresh (BF), Blood-Hematin (BH), Erosion (ERO), Ulcers (ULC), Lymphangiectasia (LYM), and Polyp (PYL) [10].*

2 Objectives

By efficiently handling the massive volume of data and the image quality issues of WCE images, an intelligent system for such modality analysis would provide an accurate identification of GI abnormalities while reducing time and cost. Upon viewing WCE images, abnormalities or lesions which are symptoms of different diseases may be revealed. Some common lesions including Blood-Fresh (BF), Blood-Hematin (BH), Foreign Body (FB), Erythema (ERY), Angiectasias (ANG), Erosion (ERO), Ulcers (ULC), Lymphangiectasia (LYM), and Polyp (PYL) [10] are shown in Fig. 1.3. Since these lesions appear in a few frames of a video and usually have a small size compared to the frame size, physicians may miss them during the examination [13]. In addition, it is a time-consuming and boring task for the physician to check a thousand frames to find pathological lesions [14]. Therefore, a computer-aided (CA) method is needed to automatically detect frames containing lesions.

In order to cope with the challenge, this thesis aims to develop a smart CA system in WCE, leading to improved diagnostic accuracy and more efficient clinical decision-making in gastroenterology. The main objective is to develop novel image-processing methodologies and intelligent algorithms based on machine-learning approaches and adapted to the specific characteristics of WCE data, in order to enhance the detection

and classification of GI abnormalities. Due to the challenge in WCE image quality, the proposed intelligent system requires a pre-processing module designed to enhance the quality of acquired images before the subsequent detection and classification stage.

In particular, with the rapid advancements in Artificial Intelligence (AI), specifically in Deep Learning (DL) approaches, Convolutional Neural Networks (CNN) is a powerful tool for image classification. Hence, in this research, the classification of GI lesions will be developed using deep convolutional networks, as they possess the capability to effectively model complex and voluminous data.

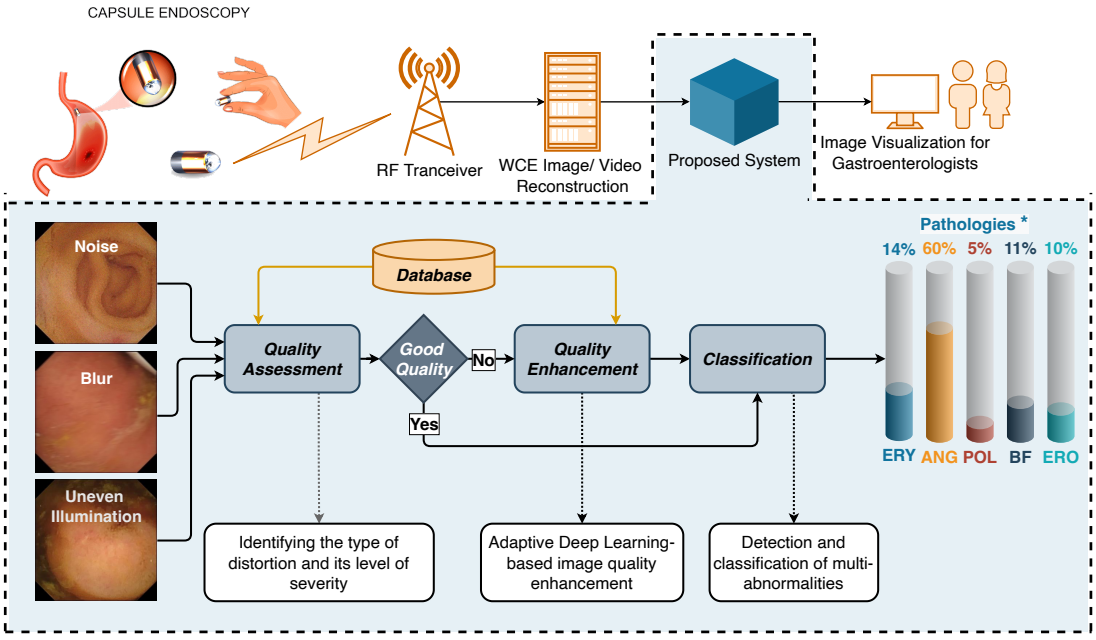


Figure 1.4: *The proposed system contains two main components of the proposed method: pre-processing and classification. The pre-processing stage involves various proposed techniques to enhance the quality of the input WCE images. The classification component employs advanced algorithms to classify the pre-processed data into distinct pathologies.*

A brief architecture of the proposed system is illustrated in Fig. 1.4. The proposed method, as depicted in Fig. 1.4, comprises two key components: pre-processing and classification. Our adaptive image enhancement methods enhance the captured frames by addressing some common distortions of WCE images such as noise, blur, and uneven illumination[15], [16]. This is achieved through a data-driven decision-making process. In order to do this, we built an image quality dataset and employed a robust quality assessment method to quantify distortions severity. By enhancing the quality of the input data, we ensure that it is in an ideal form for subsequent analysis. After pre-processing, the data is passed to the classification component, where advanced algorithms are

employed to assign appropriate class labels or pathologies to the input images. The classification algorithms exploit the extracted features and employ decision boundaries or rules to accurately classify the data. The combined implementation of pre-processing and classification within the proposed method facilitates the effective analysis and interpretation of the input data, enabling reliable decision-making in various domains such as disease diagnosis, or anomaly detection.

3 Contributions

This Ph.D. project is primarily aimed at making a significant contribution to the field of medical imaging by developing an intelligent system capable of detecting, identifying, and classifying abnormalities in Wireless Capsule Endoscopy (WCE) images. The system we are designing encompasses a crucial pre-processing module, dedicated to enhancing the quality of the captured images before they are used as input for the detection and classification phases. Furthermore, we recognize the remarkable strides made in the field of artificial intelligence, particularly within the domain of Deep Learning (DL) technologies. Convolutional Neural Networks (CNNs) have emerged as a potent approach for image classification, owing to their ability to effectively model intricate representations from complex and voluminous datasets. In our research, the classification of lesions within the WCE images will heavily rely on deep convolutional networks, exploiting their capability to extract meaningful features from these intricate medical images.

To streamline the process of detecting and identifying abnormalities, we will employ a two-step approach. First, our pre-processing module will enhance the image quality, ensuring that the input data is optimized for the subsequent analysis. Following this, a classification process will be applied, aiding in the precise identification of abnormalities within the WCE images. It's essential to note that the effectiveness of our image enhancement techniques is closely tied to the ability to take into account the degree of distortion present in the original images. The severity of distortion provides critical information that will enable us to adjust the enhancement process to an appropriate level of treatment.

Understanding the level of distortion is important because it allows us to strike a balance between improving image quality and preserving the integrity of the original data. In cases of mild distortion, a lighter enhancement approach may suffice to maintain the fidelity of the image. Conversely, when dealing with more severe distortion, a more aggressive enhancement strategy might be necessary to restore essential details and

ensure accurate detection and classification of abnormalities. To accurately identify and address distortion in WCE images, it is imperative for the proposed Deep Neural Networks) to have access to a comprehensive dataset that includes information about both the type and level of distortion present in images. Additionally, employing a well-defined assessment metric is crucial for evaluating the effectiveness of image enhancement techniques.

Therefore, in this thesis, our research unfolds systematically, commencing with the introduction of a distortion dataset, followed by the development and application of a quality assessment technique. Subsequently, we present our innovative image enhancement method, culminating in the implementation of an effective classification method. This structured approach serves as the core of our research strategy, ensuring a thorough and comprehensive examination of the challenges and solutions in the field of WCE image analysis.

Our contributions related to each method are summarized in the following sub-sections, highlighting the idea and impact of each component in the context of our research.

3.1 WCE dataset creation (QVCED)

In this work, we introduce the Quality-Oriented Database for Video Capsule Endoscopy (QVCED), an innovative dataset derived from the well-established Kvasir-Capsule dataset [10]. The primary objective of QVCED is to encompass a wide spectrum of scenarios, including various pathological conditions and multiple forms of distortion, thereby prioritizing the simulation of realistic conditions encountered in clinical practice. To construct this dataset, we employ a meticulous two-stage process.

In the initial stage, we carefully select reference videos that conform to stringent quality criteria, drawing from the Kvasir-Capsule dataset [10]. These reference videos serve as the foundation for QVCED, ensuring that the dataset maintains a high standard of quality from the outset.

Subsequently, we embark on the second stage, where we deliberately subject these reference videos to a degradation process. This process involves the controlled introduction of distortion, with the level of degradation precisely controlled using the physical parameters of a distortion generation model. By doing so, we aim to emulate and capture various types and degrees of distortion that real-world WCE videos can exhibit. This comprehensive approach equips QVCED with the capacity to address a broad

range of image quality scenarios, making it an invaluable resource for research and development in the field of video capsule endoscopy.

3.2 Image quality assessment metric (IHED)

To quantitatively evaluate the quality of images affected by uneven illumination (UI), we present a novel and straightforward Non-Reference Image Quality Assessment (NR-IQA) method, which relies on analyzing local variations in the background illuminance (BI) component of images subject to UI, such as laparoscopic images or WCE images. We employ the standard deviation of the difference signal between the BI values of the original image and its histogram-equalized counterpart as a physical metric to quantify the impact of non-uniform lighting.

The core concept behind this method is inspired by our observation that the extent to which an image is affected by UI phenomenon correlates with the degree to which contrast enhancement (CE) exacerbates the difference between the BI values before and after the application of CE.

More precisely, we propose a simple yet effective NR-IQA approach that leverages the analysis of BI variations to evaluate the impact of non-uniform lighting conditions, ultimately revealing how the application of contrast enhancement amplifies these differences. This method has the potential to offer valuable insights into the quality assessment of laparoscopic images, WCE images and their suitability for various medical applications. Indeed, we utilized this metric to meticulously create a comprehensive quality-oriented dataset including various types and levels of distortions. This dataset plays an important role in evaluating how image quality affects the performance of machine learning models for robust quality assessment.

3.3 Image quality enhancement method (TCFA)

After evaluating the distortion severity, we propose a novel image enhancement method applied to WCE images. The proposed method, namely TCFA which stands for Triplet Clustering Fusion Autoencoder, effectively addresses the challenging issues of noise, blur, and uneven illumination while accounting for varying degrees of distortion severity. Our key contributions include a distortion level encoder that classifies distortion severity using the triplet loss function, a variational cross-attention module for precise uneven illumination correction, and the integration of a pre-trained network for extracting essential structural features from WCE images. Extensive experiments conducted on the Kvasir-Capsule dataset [10] demonstrate the remarkable efficacy of TCFA, showcasing

its superior performance compared to state-of-the-art methods when applied individually to distinct distortion types, our approach focuses on addressing one distortion at a time rather than simultaneously managing multiple distortion levels. This approach shows promise in significantly advancing WCE image enhancement for specific distortions.

3.4 Classification method (DWSA-SSL)

The recent development of WCE has seen a surge in interest in computer-aided and vision-based solutions, aiming to automate the detection of abnormalities within imagery. While these developments hold great promise, one of the fundamental hurdles in building an effective computer-aided diagnostic (CAD) system for WCE lies in the limited availability of labeled data. To address these complexities and reduce the labeled-data requirement, in the final stage, we propose a novel self-supervised learning method, Dilated Window-based Self-Attention Self-Supervised Learning (DWSA-SSL), tailored for distinguishing various WCE pathologies. This innovative architecture uncovers the underlying structural information within WCE images by using unlabelled WCE images. This strategic application of SSL allows our model to capture essential semantic features without the introduction of label bias. Furthermore, we enhance the image classification quality by incorporating a transformer backbone, which reduces redundant and noisy information that may lead to biased performance. The inclusion of an attention mechanism allows us to emphasize crucial regions, particularly lesions, during diagnostic predictions. Additionally, our novel S-DWSA (Shifted-Dilated Window-based Self-Attention) module generates hybrid attention embedding, drawing from both image patches within attention windows and their interactions with neighboring windows. Lastly, we address the issue of attention map bias with a soft-cosine attention approach, ensuring a more balanced and effective distribution of attention across image regions.

4 Thesis organization

The structure of this thesis is organized as follows:

- **Chapter 2:** In this chapter, we present the state-of-the-art relative to image quality assessment metrics, image quality enhancement, and classification methods, especially the existing works that inspired our research during my thesis.
- **Chapter 3:** We propose a novel WCE distortion dataset, called the Quality-Oriented Database for Video Capsule Endoscopy (QVCED) which serves as the

primary crucial resource for evaluating the quality of Wireless Capsule Endoscopy (WCE) images and videos.

- **Chapter 4:** We propose a No-Reference IQA metric for Uneven Illumination Assessment on Laparoscopy and WCE Images, namely Illumination Histogram Equalization Difference (IHED). Subsequently, we extend its application to our proposed QVCED dataset to evaluate its effectiveness on WCE images.

- **Chapter 5:** We introduce a novel method, called **T**riplet **C**lustering **F**usion **A**utoencoder (TCFA), for enhancing the quality of WCE images which not only deals with three types of degradations including noise, blur, and uneven illumination but also considers different levels of distortion severity.

- **Chapter 6:** We propose a novel Dilated Window-based Self-Attention Self-Supervised Learning (DWSA-SSL) method to distinguish among various WCE pathologies and deal with these labeled-data requirement difficulties.

- **Chapter 7:** We provide a summary of the work presented in the previous chapters, the contributions to knowledge already achieved in this research, and the directions for the future work.

5 Publications

Journal papers:

i. T. -S. Nguyen, J. Chaussard, M. Luong, A. Beghdadi, H. Zaag and T. Le-Tien *"Dilated Window-based Self-Attention Self-Supervised Learning for Classification in Wireless Capsule Endoscopy,"* IEEE Transactions on Medical Imaging (Under preparation for submission).

ii. T. -S. Nguyen, M. Luong, J. Chaussard, A. Beghdadi, H. Zaag and T. Le-Tien *"TCFA: Triplet Clustering Fusion Autoencoder for Quality Enhancement of Wireless Capsule Endoscopy Images,"* IEEE Transactions on Medical Imaging (Under preparation for submission).

International conference papers:

i. T. -S. Nguyen, M. Luong, J. Chaussard, A. Beghdadi, H. Zaag, and T. Le-Tien, *"A Quality-Oriented Database for Video Capsule Endoscopy,"* 11th European Workshop on Visual Information Processing (EUVIP), Norway, 2023 (Winner of Three-Minute Thesis (3MT) Competition Award).

ii. T. -S. Nguyen, J. Chaussard, M. Luong, H. Zaag and A. Beghdadi, *"A No-Reference Measure for Uneven Illumination Assessment on Laparoscopic Images,"* 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 4103-4107, doi: 10.1109/ICIP46576.2022.9897302.

CHAPTER 2

Related works

Chapter content

1 Introduction 15

2 Image definition 15

3 Distortion models 16

3.1 Additive Gaussian noise model 16

3.2 Blur model 17

3.2.1 Motion blur model 17

3.2.2 Defocus blur model 18

3.3 Uneven illumination model 18

4 The State-of-the-art (SOTA) approaches 20

4.1 Introduction 20

4.2 The SOTA Image Quality Assessment (IQA) metrics 21

4.2.1 Introduction 21

4.2.2 Full-Reference IQA (FR-IQA) 21

4.2.2.1 Mathematical metrics 21

4.2.2.2 Human Visual System-based metrics 22

4.2.2.3 Other metrics 24

4.2.3 Reduced-Reference IQA (RR-IQA) 25

4.2.4 No-Reference IQA (NR-IQA) 26

4.2.4.1 Distortion-specific NR-IQA methods 27

4.2.4.2 General-purpose NR-IQA methods 29
4.2.5 metrics 34
4.2.5.1 PLCC(Pearson Linear Correlation Coefficient) 34
4.2.5.2 SROCC(Spearman Rank-Ordered Correlation Coefficient) 34
4.2.5.3 KROCC(Kendall Rank Order Correlation Coeffcient) 34
4.2.5.4 RMSE(Root Mean Squared Error) 35
4.2.6 Discussion 35
4.3 The SOTA image quality enhancement methods 35
4.3.1 Introduction 35
4.3.2 The SOTA image denoising methods 36
4.3.2.1 Classical methods 36
4.3.2.2 Dictionary learning-based methods 39
4.3.2.3 Deep-learning-based methods 40
4.3.3 The SOTA image deblurring methods 42
4.3.3.1 Classical methods 42
4.3.3.2 Deep-learning-based methods 43
4.3.4 The SOTA uneven illumination correction methods 44
4.3.4.1 Classical methods 44
4.3.4.2 Deep-learning-based methods 46
4.3.5 Discussion 47
4.4 The SOTA WCE image classification methods 48
4.4.1 Introduction 48
4.4.2 Mono-pathology classification 48
4.4.3 Multi-pathology classification 50
4.4.4 Discussion 52
5 Conclusion 52

1 Introduction

In the present day, as medical imaging is widely employed in healthcare, especially for pathological diagnosis and assessment, ensuring the high quality of medical images has become a significant concern. To attain the most accurate diagnoses, it is crucial that medical images are clear, and free of disturbing factors such as noise and other artifacts. As progress continues, improving the resolution and quality of medical images, the persistence of distortion remains a prevalent concern in the field of medical imaging. Typically, medical images are prone to various types of noises introduced during various stages such as acquisition, transmission, storage, and display. The task of image quality enhancement holds paramount importance because distortion has the risk of resulting in inaccurate diagnoses.

The primary objective of this chapter is to provide a concise overview of recent advancements in methods for assessing image quality, enhancing image quality, and applying classification techniques. In the first part of this chapter, Section 2, we introduce the fundamental definitions of continuous and discrete images, which will serve as the foundation for the subsequent discussions in this thesis. Next, Section 3 delves into various distortion models commonly encountered in WCE imaging modality. In Section 4, we explore several established image quality metrics that are utilized to assess the effectiveness of image processing techniques and evaluate the severity of distortions. Besides, some state-of-the-art approaches to image enhancement are highlighted. Additionally, Section 4 also provides an overview of the WCE image classification methods.

2 Image definition

We start by expressing a mathematical representation of an image in a discrete and continuous setting.

Definition 2.1 (Continuous Image [17], Def. 3.1)**.** *Let* $\Omega \in \mathbb{R}^d (d \in \mathbb{N})$ *be the image spatial domain. A function* $\mathbf{u} : \Omega \to \mathbb{R}$ *is called a d-dimensional image if the following conditions are fulfilled,*

i. $\mathbf{u}$ *has a compact support, if* Ω *is not bounded,*

ii. $0 \leq \mathbf{u}(x) \leq \infty$ *for all* $x \in \mathbb{R}$

iii. $\int_{\Omega} \mathbf{u}(x)dx < \infty$

This representation of images offers a simple basis for the development and analysis of mathematical methods. However, it's essential to note that it is purely an abstraction and cannot be practically implemented on any computer system. Consequently, we will focus on digital images, which are more aligned with the demands of practical applications.

Definition 2.2 (Discrete Image)**.** *Let* $\Omega \in \mathbb{N}^d (d \in \mathbb{N})$ *be a finite subset of* $\mathbb{N}^2$*. A digital image is an application* $\mathbf{u} : \Omega \to \mathbb{A}$ *such that* $\mathbb{A}$ *is a finite subset of* $\mathbb{R}^k$*, with* $k \in \mathbb{N}$*.*

In the following, unless specified otherwise, we will focus on the image such that $\mathbb{A} \in \mathbb{N}$.

In the next subsection, we list some models of common distortions found in WCE images, including additive Gaussian noise, blur, and uneven illumination.

3 Distortion models

In the domain of image processing and computer vision, an accurate modeling of distortions is useful and indispensable for various applications including image enhancement and restoration techniques. Distortions in images can originate from numerous sources, such as sensor limitations, atmospheric conditions, or transmission artifacts, and understanding and characterizing these distortions are crucial steps toward achieving robust and reliable image analysis. In this context, the following sub-section presents some mathematically tractable models that aim to characterize and quantify some common degradations in WCE such as noise, blur which can have two forms, namely motion blur and defocus blur, and uneven illumination distortion which occurs if the view field is not evenly illuminated [18]. The more efficient the degradation model is, the better we can address the impact of such degradation, by enhancing image quality, leading to improved accuracy in subsequent computer-vision tasks such as classification.

3.1 Additive Gaussian noise model

WCE cameras equipped with a limited aperture size and compact sensors featuring a constrained dynamic range often introduce noise in the captured frames [19]. In particular, the widely accepted standard noise model in WCE images is the additive white Gaussian noise [20]. Let us denote by $\mathbf{f} : \Omega \to \mathbb{R}$ with $\Omega \in \mathbb{N}^2$ the observed image

(noisy image) and $\mathbf{u}$ the ground truth image (noise-free image). As f is corrupted by random noise, each pixel value $\mathbf{f}(x)$ of $\mathbf{f}$ is the realization of a random variable F_x. We will denote by $F = (F_x)$ the random vector consisting of independent random variables F_x and $p_F(\cdot)$ the probability density function (PDF) of F. Usually, it is assumed that the random variables F_x are statistically pairwise independent and identically distributed. Then, the conditional probability density function, $p_F(\mathbf{f} \mid \mathbf{u})$, can be written as:

$$p_F(\mathbf{f} \mid \mathbf{u}) = \prod_{x \in \Omega} p_{F_x}(\mathbf{f}(x) \mid \mathbf{u}(x)) \tag{2.1}$$

An acquired image corrupted by additive Gaussian noise can be mathematically represented as:

$$\mathbf{f}(x,y) = \mathbf{u}(x,y) + \mathbf{N}(x,y), \tag{2.2}$$

where $\mathbf{f}(x,y)$ represents the pixel value at coordinates (x,y) in the noisy image, $\mathbf{u}(x,y)$ represents the original pixel value in the clean image, and $\mathbf{N}(x,y) \sim \mathcal{N}(0,\sigma_n^2)$ represents the random noise value at coordinates (x,y) following a Gaussian distribution with standard deviation σ_n.

3.2 Blur model

Capsule cameras often feature wide-angle lenses that can only capture a limited range of clear images. Blurriness can occur when the camera moves rapidly, operates at a low frame rate, or when the lens is not correctly focused. These problems can lead to two distinct types of blurriness: one caused by fast movement, namely motion blur, and the other due to improper focus, namely defocus blur.

3.2.1 Motion blur model

The degradation of blurry images can be approximated as a linear degradation system. The resultant blurring in the image, caused by the relative movement between the camera and the scene, can be represented through a two-dimensional convolution model within the linear spatial domain, and this can be formed as follows:

$$\mathbf{f} = \mathbf{u} \circledast h, \tag{2.3}$$

where $\mathbf{f}$, and $\mathbf{u}$ represent the blurred and original image respectively, and h is the motion PSF (Point Spread Function) function or blur kernel. The notation $\circledast$ represents the convolution operator.

From (2.3), it is evident that the main goal in deblurring is to estimate the PSF

function h. Assuming the scene object moves uniformly compared to the camera, we can conclude that the gray colors of points in the blurry picture are connected to the gray colors of nearby points in the original picture. This relationship allows us to define the PSF for motion-induced blurring as follows [21]:

$$h(x,y) = \begin{cases} 1, & \text{if } 0 \leq |x| \leq l\cos\theta, |y| = l\sin\theta \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

where l is the length of blur, and θ represents the angle of blur.

3.2.2 Defocus blur model

The defocus blur model resembles that of motion blur but employs a different filter. It begins by generating a Gaussian Kernel $G_0(x,y)$ [22], followed by applying Gaussian Blur through the convolution of the image with this normalized box filter.

$$G_0(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{\frac{-(x-\mu_x)^2}{2\sigma_x^2} + \frac{-(y-\mu_y)^2}{2\sigma_y^2}}, \tag{2.5}$$

where μ is the mean (the peak) and σ^2 represents the variance (per each of the variables x and y). In this work, we assume that the standard deviation in each direction is the same ($\sigma_x = \sigma_y = \sigma$), and the center of the Gaussian distribution is at the origin ($\mu_x = \mu_y = 0$) , we can simplify the 2D Gaussian blur kernel formula as follows:

$$G_0(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{2.6}$$

For the normal distribution, the values less than one standard deviation away from the mean encompass 68.27% of the data; while two standard deviations from the mean account for 95.45% of data; and three standard deviations account for 99.73%. Therefore, the rule of thumb for Gaussian filter design is to set the filter half-width to about $3 * \sigma$ (standard deviation) in each direction. Because the kernel size has been an odd number, so the kernel size is:

$$w = 2 \times \lceil 3\sigma \rceil + 1 \tag{2.7}$$

3.3 Uneven illumination model

Due to the complex structure of the GI tract, the image captured from the wireless capsule suffers from non-uniform illumination. Fig. 2.1b shows us the field of view (FOV) and illumination map of a WCE example. This image suffers from uneven

illumination with a non-uniform distribution. As we can see in Fig. 2.1, when the

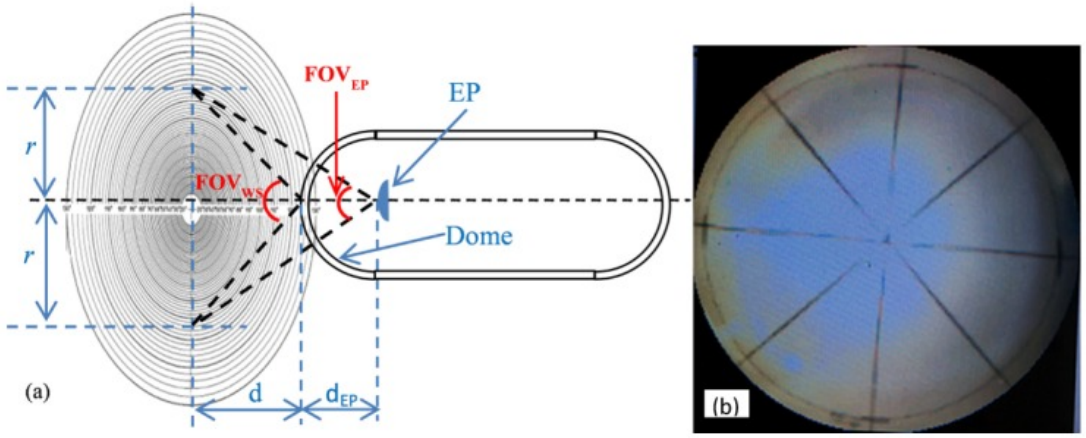
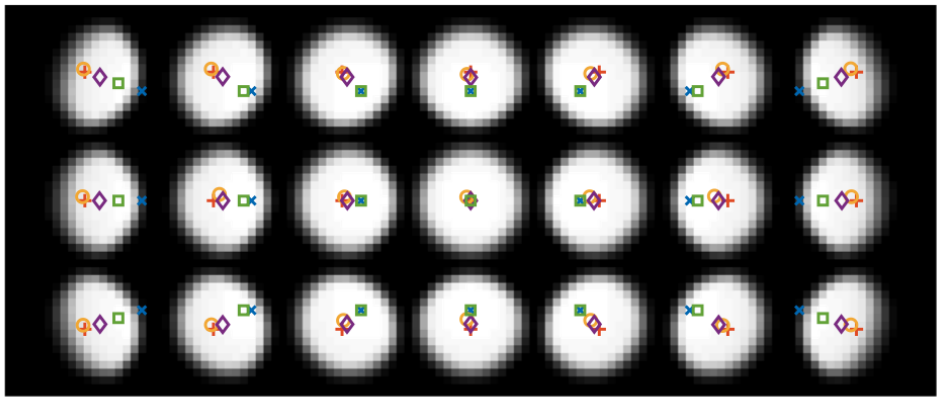


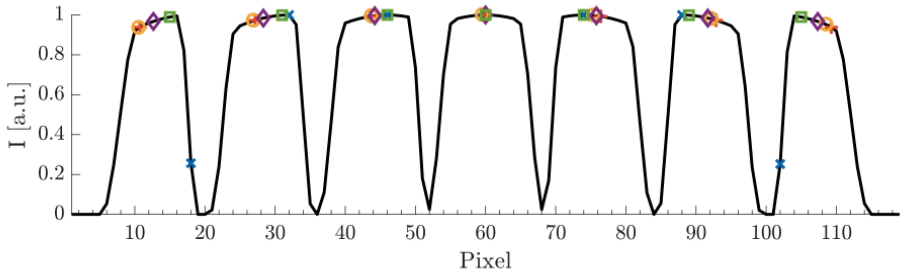
Figure 2.1: *Field of View (FOV) measurement of a capsule endoscope: (a) illustration of FOV_{WS} (window surface) and FOV_{EP} (Entrance Pupil), (b) a low-resolution image ($4cm \times 4cm$) on a capsule recorder screen [23].*

light source is not evenly distributed across the entire FOV, it can lead to reduced illumination towards the edges of the image, causing a vignette effect as a specific type of uneven illumination. This is particularly noticeable when the illumination source is not centered or when there are obstacles or irregularities in the optical path.

To analyze the effect of the direction of the light source on a captured uneven illumination effect, Lois et al. [24] established a dedicated microscopy system. it should be noted that both microscopy and WCE utilize optical systems to capture images. While the scale and complexity differ, the fundamental principles remain the same - light passes through a lens system onto an image sensor, where the quality of the image is influenced by the distribution of light. Fig. 2.2 shows us an example result of uneven illumination when the direction of the light source is not straight. As we can see in Fig. 2.2, if the light source is not straight to the object, the center line of the illumination will appear as an approximately log-normal distribution. Based on the experiment result, we could assume that the uneven illumination in WCE will have a 2D hybrid normal-log-normal distribution shape. The specificities of the distortion simulation will be elaborated upon in Chapter 3.



(a) White Image



(b) Intensity profile of the center row

Figure 2.2: *(a) Light-shading images when the light source is not directly aligned with the object [24], consider a scenario where the light source is positioned on the surface of a spherical object, with the object in the center of this sphere. (b) The highest intensity point (marked in green) within a sub-image corresponds to the point where the microlens center (marked in blue) is orthogonally projected. The actual camera's central view is marked as the ground truth point (marked in red). This ground truth point can also be alternatively represented using a mathematical model as the reference point (marked in yellow), and the weighted average point (marked in violet) includes a weighted average of pixel intensities within the sub-image.*

4 The State-of-the-art (SOTA) approaches

4.1 Introduction

The main objective of this thesis is to develop a smart system with the capability to detect, identify, and classify abnormalities found in WCE images. An interesting and essential element of the proposed system is a pre-processing module dedicated to enhancing the quality of the acquired images before they are utilized in the detection and classification stages. Moreover, for a better performance, we propose to design a learning-based image enhancement that necessitates the knowledge of the distortion level, the latter can be obtained through the use of IQA metrics. By integrating distortion

level information in the pre-processing module, the overall system aims to optimize the image quality enhancement, thereby enabling more precise and reliable detection and classification of abnormalities in WCE images. In this context, we first make a comprehensive overview of the state-of-the-art of three main groups including SOTA IQA metrics, SOTA image quality enhancement, and SOTA image classification on WCE images.

4.2 The SOTA Image Quality Assessment (IQA) metrics

4.2.1 Introduction

Typically, the performance of the image processing methods is often evaluated in both objective measures and subjective quality assessment protocols. The objective measure is determined through the image quality metrics. While subjective assessment of image quality is considered highly reliable and serves as a benchmark for objective quality metrics, it is a complex and time-consuming process that necessitates a controlled environment adhering to ITU standard requirements [25]. Moreover, subjective evaluation is not feasible in real-time applications where quick responses and decisions are required. Consequently, a multitude of objective image quality metrics have been developed. Objective Image Quality Assessment (IQA) methods can be categorized into three main types: Full-Reference IQA (FR-IQA) [26], Reduced-Reference IQA (RR-IQA) [27], and No-Reference or Blind IQA (NR-IQA) [28]–[30].

4.2.2 Full-Reference IQA (FR-IQA)

A Full-Reference Image Quality Assessment (FR-IQA) metric is a method used to evaluate the perceptual quality of an image to a reference one (a high-quality image). It requires access to both the distorted image and the corresponding reference image for a direct comparison. FR-IQA metrics aim to quantitatively assess the similarity between the distorted and reference images based on various visual characteristics, such as structural information, color fidelity, texture details, and perceptual artifacts. The image similarity measures are grouped according to their strategies including mathematical metrics, Human Visual System (HVS) based metrics, and others.

4.2.2.1 Mathematical metrics

In the mathematical approach, the image is regarded as a 2D signal, and the dissimilarity (or similarity) between the reference and the distorted images is calculated as a distortion (or quality) measure. Minkowski metric, for example, calculate the E_γ distance between

the reference image $\mathbf{x}$ and distorted image $\mathbf{y}$ by:

$$E_\gamma = \left(\frac{1}{N} \sum_{i=1}^{N} |\mathbf{x}_i - \mathbf{y}_i|^\gamma \right)^{1/\gamma}, \tag{2.8}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the i-th samples in image $\mathbf{X}$ and $\mathbf{Y}$ respectively, N is the number of image samples, and $\gamma \in [1 \ldots \infty)$. For $\gamma = 2$, one obtains the well-known Mean Square Error (MSE) formula if the square root is ignored:

$$E_2 = \left(\frac{1}{N} \sum_{i=1}^{N} |\mathbf{x}_i - \mathbf{y}_i|^2 \right)^{1/2} = \sqrt{MSE} \tag{2.9}$$

MSE and its variants such as peak signal-to-noise ratio (PSNR) are commonly used as objective image quality metrics. PSNR is defined as:

$$PSNR = 10 log_{10} \left(\frac{255^2}{MSE} \right) \tag{2.10}$$

The value 255 represents the maximum gray level in an 8-bit per pixel monochromatic image. Additional mathematical measures for quantifying image distortion, such as average difference, maximum difference, absolute error, Peak MSE, and others, are detailed in [31]. In the same study, Eskicioglu and Fisher also introduced correlation-based measures, including structural content (SC), normalized cross-correlation (NCC), correlation quality, and others. These measures are designed to evaluate the likeness between reference and test images.

However, it's worth noting that despite their simplicity and mathematical precision, these metrics do not exhibit a strong correlation with perceived quality measurements [32], [33]. This lack of correlation primarily arises from the fact that these metrics do not take into account the characteristics of the HVS in their models. The HVS plays a substantial role in shaping our perception of image quality.

4.2.2.2 Human Visual System-based metrics

In this approach, the error signal, which represents the disparity between the reference and test images, is evaluated through known properties of the human visual system. Several characteristics of the HVS frequently employed in image quality assessment (IQA) including contrast sensitivity function (CSF), luminance contrast sensitivity and contrast masking.

Firstly, CSF plays a role in the integration process by either acting as a filter to enhance lower spatial frequencies or by assigning different weights to subbands after

frequency decomposition. This is because the human visual system is more sensitive to lower spatial frequencies compared to higher ones, as observed by Lee in [34].

Secondly, based on the luminance contrast sensitivity of the human vision, it stands out that our ability to detect variations in brightness, referred to as luminance contrast, surpasses our capacity to precisely evaluate the actual brightness level. This observation remains consistent with Weber's law, which maintains that the relationship between the smallest noticeable change in brightness (ΔL) and the initial brightness value (L), denoted as $\Delta L/L$. $\Delta L/L$ is constant across a broad spectrum of brightness levels. In scenarios characterized by low background brightness, such as dimly lit environments, this $\Delta L/L$ ratio becomes notably more pronounced. This phenomenon is frequently termed "luminance masking" in various IQA methodologies. Essentially, it implies that in poorly illuminated environments, our ability to perceive and differentiate elements within an image diminishes due to the reduced information pertaining to luminance contrast [35].

Thirdly, contrast masking occurs when an element within an image becomes more difficult to distinguish due to the presence of nearby objects. This phenomenon plays a fundamental role in the way our eyes function. Within models utilized for IQA, contrast masking is deliberately used to replicate situations in which certain image details may become less visible or even concealed by adjacent elements in the image. These circumstances can lead to alterations in our perception of the image's quality. By encompassing the impacts of contrast masking, IQA metrics strive to provide a more precise assessment of image quality, taking into account the intricate interactions among various visual components within the image.

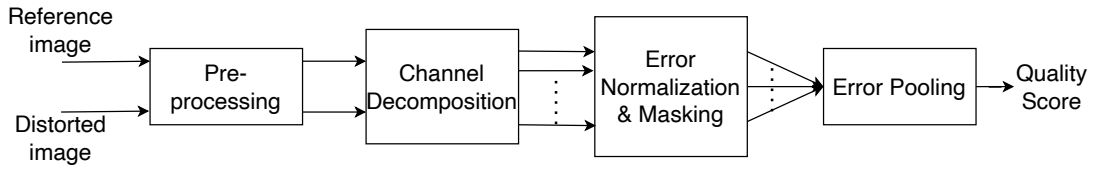


Figure 2.3: *A standard framework of the Image Quality Assessment (IQA) system based on the Human Visual System (HVS) involves a pre-processing phase preceding the channel decomposition. This pre-processing phase encompasses operations such as alignment, conversion of color spaces, and low-pass filtering through the point-spread function (PSF) to simulate the effects of eye optics.*

The HVS features, as described in Fig. 2.3, are implemented at the pre-processing and the error normalization and masking stage. The "Channel decomposition" block serves the purpose of changing the values of individual pixels in an image into separate and less

related spatial sub-components. This transformation has the potential to enhance the accuracy of quality measurement. Several transformation methods have been suggested for this purpose, including the cortex transform, steerable pyramid transform, wavelet transform, and DCT transform. The "Error normalization and masking" blocks are applied individually to each channel within a system. In the majority of models, masking is realized through a gain-control mechanism, which adjusts the weighting of the error signal in a given channel based on a spatially varying visibility threshold specific to that channel. Finally, the "Error pooling" block refers to the process of consolidating or aggregating the error signals from multiple channels into a unified distortion or quality assessment score. This technique is commonly used in various applications, such as image or video processing, to summarize the overall quality or distortion across different components or regions of the images [34].

The first effort to employ the HVS for measuring image quality was made by Mannos and Sakrison [36]. This idea was later expanded upon by various other researchers. Some notable models in this category include the "visible differences predictor" by Daly [37], Watson's DCT-based and Wavelet-based Metrics [35], the "perceptual image distortion" model by Teo and Heeger [38], and the Just-Noticeable-Distortion (JND) model by Lubin [39]. Although this approach is widely accepted, it does have some limitations. These include the nonlinearity of the HVS, as well as challenges related to detecting subtle differences beyond a certain threshold, known as the "suprathreshold problem" [40].

4.2.2.3 Other metrics

Some other metrics are proposed in a simpler manner to be consistent with visual perception. Wang et al. have introduced a novel framework for creating image quality metrics [40], [41], known as the "structural similarity" approach. In this approach, they make the assumption that the HVS is exceptionally adapted for perceiving information in natural scenes, which are typically structured. Therefore, they propose that assessing the change in structural information (between a reference and a distorted image) should provide a reliable approximation of perceived image distortion. The SSIM value between two images is determined by averaging the similarity index between corresponding local image patches, denoted as $\mathbf{x}_i$ and $\mathbf{y}_i$, both of which belong to the same position in the two images being compared.

Concretely, the local SSIM index measures the similarities of three elements of the image patches: luminance, contrast and structure as indicated in (2.11), (2.12)

and (2.13).

$$\ell(\mathbf{x}_i, \mathbf{y}_i) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}, \tag{2.11}$$

$$c(\mathbf{x}_i, \mathbf{y}_i) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}, \tag{2.12}$$

$$s(\mathbf{x}_i, \mathbf{y}_i) = \frac{2\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}, \tag{2.13}$$

where μ_x, μ_y are the means of the images patches $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively; σ_x,σ_y the standard deviations of $\mathbf{x}_i$ and $\mathbf{y}_i$, and $\sigma(\mathbf{x}_i, \mathbf{y}_i) = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{x}_i(j) - \mu_x)(\mathbf{y}_i(j) - \mu_y)$ the covariance of the two images patches $\mathbf{x}_i$ and $\mathbf{y}_i$, while c_1, c_2, c_3 are positive constant to stabilize the division with weak denominator. Then, the SSIM defined locally between two image patches is defined as a weighted combination of three comparative measures.

$$SSIM(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{M} SSIM(\mathbf{x}_i, \mathbf{y}_i) = \sum_{i=1}^{M} \ell(\mathbf{x}_i, \mathbf{y}_i)^{\alpha} \cdot c(\mathbf{x}_i, \mathbf{y}_i)^{\beta} \cdot s(\mathbf{x}_i, \mathbf{y}_i)^{\gamma}, \tag{2.14}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the reference and the test images, respectively, while $\mathbf{x}_i$ and $\mathbf{x}_i$ are the image patches at the i-th local window, and M is the number of local windows in the image. α, β, γ are positive constants.

In [42], the mutual information between the test and the reference images, is quantized to relate to visual quality. The proposed measure is called Visual Information Fidelity (VIF). The concept behind VIF is to capture and quantify the extent to which the visual information in a test image aligns with that in a reference image. By analyzing the mutual information between these two images, VIF aims to provide a measure that reflects how faithful the test image is to the reference image in terms of preserving visual information. This approach offers a valuable tool for evaluating and comparing the quality of images by considering the information content shared between them.

4.2.3 Reduced-Reference IQA (RR-IQA)

Reduced reference image quality assessment (RR-IQA) constitutes an approach that approach that extracts a minimal but appropriate set of characteristics from the reference image to evaluate the quality of a test image. These characteristics encompass elements such as frame intricacies, edges, and color attributes, functioning as partial features. To attain efficacy, an RR-IQA algorithm must conform to a set of essential criteria, meticulously delineated in Wang and Simoncelli's seminal work [43]:

- The algorithm's primary function should involve generating a succinct and efficient summary of the reference image, extracting essential information for the quality assessment process.

- The algorithm must demonstrate a robust capacity to accurately identify a spectrum of distortions inherent in the test image. Proficiency in discerning diverse distortion types facilitates a holistic evaluation of image quality.
- Additionally, the RR-IQA algorithm must correlate human perception of image quality. Its design should encompass the capture and analysis of visual attributes appropriate to human observers, thereby ensuring the relevance between algorithmic assessments and human judgements.

In the domain of communication systems, these algorithms serve as essential tools. Their main purpose is to oversee the quality of images as they traverse through communication channels, effectively identifying any distortions that might crop up during the transmission process. Through the utilization of RR-IQA algorithms, it becomes viable to measure the influence of these distortions on image quality, thereby facilitating the implementation of suitable corrective actions.

4.2.4 No-Reference IQA (NR-IQA)

NR-IQA algorithms are formulated to forecast the subjective quality of altered images by relying on human perception, all without necessitating access to the corresponding reference images, also known as ground truth images. These algorithms hold significant utility in real-world situations where obtaining or accessing the reference image proves to be prohibitively expensive or infeasible. Although humans can frequently intuitively assess the subjective quality of a distorted image even in the absence of a reference image, replicating this from a computational perspective has presented considerable challenges.

Over the years, there has been a development of numerous NR-IQA algorithms, which can be broadly classified into two categories: distortion-specific and general-purpose (universal) approaches. Distortion-specific NR-IQA algorithms utilize particular distortion models to estimate the subjective quality of an image, as mentioned in Corchs et al.'s work [44]. These algorithms have the capability to measure one or more distortions within an image, such as noise and blur, and subsequently assess its visual quality based on these parameters. These NR-IQA algorithms are adept at quantifying and evaluating various types of distortions present in an image, which may include blur, blocking, ringing, and noise. They are often referred to as application-specific NR-IQA algorithms since their applicability depends on prior knowledge of the specific distortion type.

In contrast, general-purpose NR-IQA algorithms present a more formidable task since

they aspire to function effectively across various types of distortions, devoid of any prior knowledge about the specific distortion type. This is due to practical situations where information regarding the distortion type and its statistical characteristics may be either unavailable or entirely unfamiliar. Moreover, an image's quality can be simultaneously influenced by multiple types of distortions, thereby compounding the complexity of the NR-IQA problem in terms of analysis, research, and development. Although existing NR-IQA algorithms still grapple with limitations concerning image degradation modeling and training, as noted in Beghdadi's survey [45], general-purpose NR-IQA algorithms hold substantial potential for application in uncontrolled environments, as indicated by Shahid [46], despite these formidable challenges.

After briefly introducing NR-IQA algorithms, we review in detail various approaches of distortion-specific and general-purpose NR-IQA algorithms in Section 4.2.4.1 and Section 4.2.4.2, respectively.

4.2.4.1 Distortion-specific NR-IQA methods

The fundamental goal of research in the NR-IQA field is to devise computational models that can accurately predict image quality, even without or with little available prior information about the image's contents and distortions. Typically, all NR-IQA algorithms adhere to a two-step framework, which involves extracting features and mapping them to subjective quality. However, when it comes to practical implementation, distortion-specific NR-IQA algorithms, which specialize in identifying specific distortions like blurring and noise in WCE images, tend to be simpler in terms of feature extraction and quality prediction compared to their general-purpose NR-IQA counterparts. In the following section, we will provide a concise overview of the distortion-specific NR-IQA algorithms.

Blur-oriented algorithms Blur artifacts tend to have a notable impact on main image elements, including edges and intricate details. These effects can be examined in either the spatial domain or the spatial-frequency domain [47]. NR-IQA algorithms that specifically address blur concentrate on modeling the dispersion of edges and establishing a connection between this dispersion and the perceived image quality, as illustrated in Sang et al.'s work [48]. Over time, various techniques have been developed to quantify blur artifacts. These methods encompass calculating the block kurtosis of discrete cosine transform (DCT) coefficients [49], detecting edge widths [50], evaluating image spectrum uniformity [51], scrutinizing the dominant eigenvalues of the covariance matrix [52], employing iterative thresholding of a gradient image [53], and utilizing

probabilistic approaches for blur detection [54].

To improve the consistency and effectiveness of NR-IQA algorithms focused on blur, researchers have introduced models that reflect how blurriness is perceived by the HVS. This approach seeks to align these algorithms with the characteristics and mechanisms of human perception. For instance, in [55], Sadaka et al. explored the use of saliency models to construct a blur-oriented NR-IQA algorithm. Additionally, in [56], the authors introduced a perceptual sharpness measure based on just noticeable blur (JNB). JNB considers how the HVS responds to sharpness at various contrast levels, enabling the quantification of blurriness perception. To account for the impact of noise on blur assessment, a noise-robust blur measure was proposed in [57]. This measure combines a gradient-based approach with singular value decomposition to enhance its resistance to noise interference. Furthermore, a more recent approach, as detailed in [58], employs multiplicative multi-resolution decomposition to analyze blurring effects in an innovative manner. These advancements in modeling blurring, incorporating HVS characteristics, and applying novel techniques have significantly contributed to the development of more effective blur-oriented NR-IQA algorithms.

Noise-oriented algorithms Accurate noise statistics estimation holds significant importance for the precise assessment of image quality in the presence of noise. This is because noise estimation can be utilized for evaluating image quality, as demonstrated by Gabarda's and Zhai's works [59], [60]. Within the literature, various noise estimation approaches have been developed, including the filter-based approach [61], the patch-based approach [62], and a combination of both these approaches [63].

Recent advancements in noise estimation algorithms have developed. For instance, an IQA algorithm proposed in [64] utilizes a block-based homogeneity measure for estimating noise statistics like variance within the image. This algorithm also takes into account the visual masking effect of the HVS. Another challenge in noise estimation is the selection of an appropriate block size, as it directly affects the accuracy of locally derived noise statistics. To address this challenge, researchers introduced a deformable ant colony optimization (DACO) algorithm in [65]. This algorithm dynamically adjusts the kernel size for image block selection, allowing for improved noise estimation. Deng et al. [66] examine how well the kurtosis of image wavelet coefficients, when used as a feature for a feedforward NN, can be employed to teach the NN to convert kurtosis values into scores related to perceptual quality. These recent developments in noise estimation algorithms contribute to the more accurate assessment of image quality for noisy images by effectively capturing and analyzing the underlying noise characteristics.

Multiple distortions-oriented algorithms In addressing the complexity arising from the simultaneous presence of various types of distortions, multiple approaches have been highlighted [67]–[69]. For instance, Li introduced an NR-IQA algorithm [69] that integrates metrics related to three distinct image distortions: ringing, noise, and blocking. Its primary goal revolves around the identification and quantification of the impact these distortions have on image quality. Another approach, presented by Cohen and Yitzhaky [70], focuses specifically on noise and blur, aiming to discern and measure their effects on image quality. This particular algorithm operates by meticulously examining shared statistical characteristics derived from the power spectra of the images. Furthermore, Sazzad et al. proposed an NR-IQA algorithm [71] that remains unrestricted by predefined artifacts commonly associated with JPEG2000 images. Its purpose is to evaluate image quality without being constrained by the presence of specific artifacts.

These efforts collectively highlight the evolution of NR-IQA algorithms geared toward addressing the simultaneous influence of diverse distortion types, thereby providing comprehensive evaluations of image quality in a variety of scenarios.

4.2.4.2 General-purpose NR-IQA methods

In the field of NR-IQA, there has been a shift towards adopting machine learning-based approaches, primarily driven by the demand for applications based on machine learning. As a result, most existing NR-IQA algorithms follow a two-step framework [72], consisting of feature extraction and quality prediction. This framework aims to learn the human evaluation process using human-rated scores available in benchmark databases. The objective is to develop algorithms that can automatically predict the visual quality of images across different types of distortions and image contents. Similar to distortion-specific NR-IQA algorithms, when developing a robust general-purpose NR-IQA algorithm, two major challenges arise: features extraction and quality prediction. These challenges are more complex in the context of general-purpose algorithms and require in-depth exploration and discussion. In the following subsection, we will examine these issues in detail and discuss approaches for addressing them effectively.

Features Extraction The natural world represents only a small portion of all potential scenes, and it's far from random in its appearance. Instead, it demonstrates consistent statistical characteristics known as Natural Scene Statistics (NSS), which have drawn significant attention from researchers in their investigation of statistical features that correlate well with human judgements of quality. In the domain of NR-IQA, the attributes used to evaluate image quality primarily originate from statistical properties

of natural images, whether in the spatial or transform domain.

In [73], Mittal et al. introduced an NR-IQA algorithm named the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE). This algorithm functions within the spatial domain and relies on statistics of locally normalized luminance coefficients to measure possible losses of "naturalness" in the image affected by distortions. The core principle is that natural images demonstrate specific, consistent statistical properties, which are quantifiably altered by the presence of distortions. BRISQUE employs a framework that uses a Generalized Gaussian Distribution (GGD) to model the mean-subtracted contrast-normalized (MSCN) luminance coefficients distributions, assuming that characteristic statistical properties of MSCN coefficients are modified due to the presence of distortion. It evaluates an image's naturalness by examining the parameters of the GGD model. It has been proved that the GGD model can effectively capture a wider range of statistics related to distorted images. The GGD model (with zero mean) is expressed as follows:

$$f(x;\alpha,\sigma^2) = \frac{\alpha}{2\beta\Gamma(1/\alpha)} exp\left(-\left(\frac{|x|}{\beta}\right)^{\alpha}\right), \tag{2.15}$$

where $\beta = \sigma\sqrt{\frac{\Gamma(1/\alpha)}{\Gamma(1/3\alpha)}}$ and $\Gamma(\cdot)$ is the gamma function with:

$$\Gamma(a) = \int_0^{\infty} t^{a-1}e^{-t}\,\mathrm{d}t,\ a > 0. \tag{2.16}$$

Within the BRISQUE algorithm, the parameters α and σ control the shape and variance of the distribution, respectively. The estimation of parameters α and σ is carried out through the moment-matching-based approach [74]. Subsequently, these estimated parameters find utility in quantifying the naturalness of an image.

In terms of modeling the empirical distributions of pairwise products among adjacent MSCN coefficients across various orientations (horizontal, vertical, main-diagonal, and secondary-diagonal), the BRISQUE algorithm relies on an asymmetric generalized Gaussian distribution (AGGD) model. The specific mathematical formulation of the AGGD model is described within the BRISQUE algorithm as:

$$f(x;v,\sigma_l^2,\sigma_r^2) = \begin{cases} \frac{v}{(\beta_j+\beta_r)\Gamma(1/v)} exp\left(-\left(\frac{-x}{\beta_l}\right)^{v}\right), & \text{if } x < 0, \\ \frac{v}{(\beta_j+\beta_r)\Gamma(1/v)} exp\left(-\left(\frac{x}{\beta_r}\right)^{v}\right), & \text{if } x \geq 0, \end{cases} \tag{2.17}$$

where the distribution's shape is controlled by the shape parameter v, and the spread on each side of the model is governed by scale parameters σ_l^2 and σ_r^2. The scale parameters β_l and β_r are derived as $\beta_l = \sigma_l\sqrt{\frac{\Gamma(1/v)}{\Gamma(3/v)}}$ and $\beta_r = \sigma_r\sqrt{\frac{\Gamma(1/v)}{\Gamma(3/v)}}$. It is worth noting that

BRISQUE distinguishes itself from other NR-IQA algorithms by not requiring any transformation into a different coordinate frame.

In another work, Liu et al. proposed CurveletQA as a general-purpose NR-IQA algorithm in [75]. CurveletQA operates in the curvelet domain and extracts a set of features based on Natural Scene Statistics (NSS). These features exhibit strong correlations with human perception, making CurveletQA unique in its approach. The mathematical definition of the discrete curvelet transform for a 2D function $f(t_1,t_2)$ is as follows:

$$\theta(j,l,k) = \sum_{0\leq t_1,t_2\leq 1} f[t_1,t_2]\overline{\varphi(j,l,k)[t_1,t_2]}, \tag{2.18}$$

where $\varphi(i,j,k)$ represents a curvelet with scale j, positioned at index k with angle index l, while t_1 and t_2 represent spatial domain coordinates. To enhance the effective capture of distribution characteristics of coefficients with larger amplitudes, the algorithm computes the empirical probability distribution function (PDF) of the logarithm of the magnitude of curvelet coefficients at scale j as follows:

$$h_j(x) = PDF(log_{10}(|\theta_j|)), \tag{2.19}$$

where θ_j is the set of curvelet coefficients at scale j. Then, an AGGD model is used to fit the curve of $h_j(x)$ and the fitting parameters are extracted as the features.

In [76], Saad et al. presented BLIINDS, an NR-IQA algorithm that relies on DCT statistics to make predictions about image quality. BLIINDS extracts four distinct features from the DCT domain, considering two different spatial scales. It operates at the level of local image patches, with a particular focus on DCT-based contrast and DCT-based structural characteristics. The computation of the DCT-based contrast feature involves averaging the ratio between the magnitudes of non-DC DCT coefficients within the local patch, normalized by the DC coefficient of that patch.

In a different work, Moorthy and Bovik, in [77], introduced the Blind Image Quality Index (BIQI), which estimates image quality by leveraging statistical attributes obtained through the Discrete Wavelet Transform (DWT). BIQI operates across three spatial scales and three orientations and models the sub-band coefficients derived from the DWT using a Generalized Gaussian Distribution (GGD). From this distribution, two parameters are estimated and utilized as features. The resulting 18-dimensional feature vector serves the purpose of characterizing the distortions present within the image. In their subsequent research [78], they introduced the DIIVINE algorithm, an advancement building upon the foundation laid by BIQI. DIIVINE employs a steerable pyramid transform characterized by two scales and six orientations to extract its distinctive

features. Similar to BIQI, DIIVINE makes use of a GGD model for capturing the statistical properties inherent in sub-band coefficients. Furthermore, it integrates both a joint statistical model and a structural similarity index to encode local dependencies across various dimensions, including scales, orientations, and spaces. The parameters emerging from these models, culminating in a comprehensive set of 88 features, are skillfully employed in the estimation of image quality.

In [28], Mittal et al. introduced the NIQE algorithm. NIQE's operational principle is rooted in the observation that natural images exhibit consistent statistical properties that can be leveraged to distinguish between high-quality and distorted images. NIQE calculates a quality score by scrutinizing diverse statistical features derived from the image. These features encompass aspects such as image sharpness, naturalness, noise, and contrast. The algorithm employs a learning-based approach, where a model is trained using a dataset of images with known quality scores to estimate image quality.

In another work [79], Zhang et al. proposed a novel and effective NR-IQA algorithm known as the Integrated Local Natural Image Quality Evaluator (IL-NIQE). To more comprehensively capture structural distortions and the nuances of contrast distortion, they incorporated a quality-aware gradient statistics feature. The gradient magnitudes (GMs) of natural images are amenable to modeling through a Weibull distribution, as follows:

$$p(x;a,b) = \begin{cases} \frac{a}{(b^a)} x^{a-1} exp\left(-\left(\frac{x}{b}\right)^a\right), & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \tag{2.20}$$

where parameters a and b correspond to GM and local contrast distributions of an image, respectively.

Quality prediction Within the domain of general-purpose NR-IQA algorithms, the primary goal is to establish a functional link between the quality-aware attributes derived from an image and the corresponding evaluations in terms of Mean Opinion Scores (MOS) or Differential Mean Opinion Scores (DMOS). These scores are pivotal for evaluating the performance of an IQA algorithm by its correlation with human subjective judgements of quality. Typically, this work is carried out in the framework of machine learning, as explored in various studies [80]–[84].

Certainly, neural networks (NN) have attracted considerable attention as potent mathematical tools in a range of pattern recognition applications [85], [86]. The interest in NN is increased by their ability to approximate the functional relationship between a known set of input and output data, enabling them to construct precise

models for a diverse set of tasks. In [81], Li et al. proposed an NR-IQA algorithm focused on a generalized regression neural network (GRNN). This particular algorithm estimates image quality by approximating the functional connection between quality-aware attributes, encompassing metrics such as the mean value and entropy of the phase congruency image, in addition to the entropy and gradient of the distorted image, and subjective quality ratings.

In another approach [87], the challenge of estimating image quality without reliance on a reference image is addressed using a Growing and Pruning Radial Basis Function (GAP-RBF) network. This strategy transforms the estimation task into one of function approximation, wherein the previously unknown relationship between the features and Mean Opinion Score (MOS) values is acquired during the training phase of the GAP-RBF network.

Moreover, Ghadiyaram and Bovik introduced a pioneering NR-IQA algorithm [88], based on Natural Scene Statistics (NSS). This algorithm exploits a deep belief network to explore potential feature representations and trains a quality predictor. It encompasses an unsupervised pre-training phase followed by a supervised fine-tuning stage, distinguishing itself as one of the earliest NR-IQA algorithms capable of predicting perceptual image quality even when confronted with intricate combinations of distortions.

In [89], Li et al. proposed the Shearlet and Stacked Autoencoders-based No-Reference Image Quality Assessment (SESANIA) algorithm, a versatile NR-IQA methodology that leverages a deep neural network. SESANIA extracts fundamental features using a multi-scale directional transformation known as the shearlet transform and characterizes the characteristics of natural and distorted images through the summation of sub-band coefficient amplitudes. Stacked autoencoders are employed to refine these features and render them more distinct.

Non-Reference Image Quality Assessment (NR-IQA) algorithms offer a means to evaluate the quality of an image without relying on a reference image or its features. This task presents a notably challenging problem when compared to other image quality assessment tasks. The absence of a reference image necessitates modeling the statistical properties of a reference image, understanding the characteristics of the HVS, and discerning the impact of distortions on image statistics in an unsupervised manner. Furthermore, evaluating the effectiveness of a quality measure with a specific distorted image in the absence of a reference image poses significant difficulty. Different NR-IQA algorithms yield varying quality scores. Consequently, to enable the comparison of different NR-IQA algorithms, the establishment of a common benchmarking system

becomes imperative. In the following subsection, several performance evaluation metrics for benchmarking are presented.

4.2.5 metrics

To evaluate the NR-IQA metrics, some common correlation evaluation metrics including PLCC, SROCC, KROCC and RMSE are used, as defined below.

4.2.5.1 PLCC(Pearson Linear Correlation Coefficient)

$$PLCC(Q_{est}, Q_{sub}) = \frac{Cov(Q_{est}, Q_{sub})}{\sigma(Q_{est})\sigma(Q_{sub})}, \tag{2.21}$$

where Q_{est} and Q_{sub} represent the sets of predicted and actual subjective scores, respectively, $Cov(.)$ signifies the covariance between Q_{est} and Q_{sub}, while $\sigma(.)$ denotes the standard deviation. The term PLCC serves as an indicator of the correlation between the algorithm's output and the subjective assessments made by human observers. It quantifies the accuracy of the algorithm's predictions and falls within the range of "-1" to "+1." A value close to "+1" signifies a strong positive correlation between the two variables, while a value close to "-1" indicates a strong negative correlation. Conversely, a very low or zero value suggests a lack of correlation between the two variables.

4.2.5.2 SROCC(Spearman Rank-Ordered Correlation Coefficient)

$$SROCC(Q_{est}, Q_{sub}) = 1 - \frac{6\sum d_i^2}{N(N^2-1)}, \tag{2.22}$$

where d_i represents the difference in ranks between the i-th samples of Q_{est} and Q_{sub}, and m denotes the number of images in the database. The SROCC is mainly employed to assess the monotonic relationship between two sample sets. Its value can span from "-1" to "+1," with the same interpretational conventions as Pearson's correlation coefficient. N is the number of images. When there are tied observations, they are assigned the same average rank. For instance, if three observations are tied for the third smallest value, their ranks would be 3, 4, and 5. The average rank for these three is 4, and each of the three observations would receive a rank of 4 [90]. Ties in data do not pose any issues when calculating the Spearman correlation coefficient using the Pearson formula with ranks.

4.2.5.3 KROCC(Kendall Rank Order Correlation Coeffcient)

$$KROCC = \frac{2(n_c - n_d)}{N(N-1)}, \tag{2.23}$$

where n_c is the number of concordant element pairs in the dataset; n_d is the number of discordant element pairs in the dataset, and N is the number of images. KROCC is also a good measure of the monotonicity of the two sample sets.

4.2.5.4 RMSE(Root Mean Squared Error)

$$RMSE = \left[\frac{1}{N} \sum_{i=1}^{N} (MOS_i - Pred_i)^2 \right], \tag{2.24}$$

where MOS_i is the subjective MOS value, $Pred_i$ is the quality prediction score and N is the number of images. RMSE is a direct measure of the absolute deviation between a person's subjective score and the algorithm's predicted score. Its value lies between "0" and "+1" where the value close to "+1" indicates that the two variables have a positive correlation. Zero value implies that the two variables are not correlated.

4.2.6 Discussion

It is important to acknowledge that numerous metrics have been developed to estimate objective image quality, considering various types of distortions such as additive noise, blurring, or compression artifacts [58], [91]. However, the existing NR metrics for evaluating uneven illumination, particularly in the context of medical imaging, are limited in number [92], [93]. This gap in the literature highlights the need for a new metric specifically designed to assess uneven illumination. In order to meet this demand, we propose the development of a novel metric that can effectively evaluate and quantify the impact of uneven illumination in medical images such as WCE modality. By doing so, we aim to provide a comprehensive tool to assess and improve the quality of medical images affected by uneven illumination, enhancing their diagnostic potential and overall utility in clinical practice.

4.3 The SOTA image quality enhancement methods

4.3.1 Introduction

WCE image acquisition often results in degradation due to camera limitations and environmental factors, such as a narrow aperture and small sensors producing noise and uneven illumination, or unstable conditions causing blurry images. These distortions can significantly hinder the experiences of clinicians and potentially lead to erroneous diagnoses. Consequently, the development of image enhancement methods for improving the quality of WCE images becomes imperative. These techniques aim to mitigate the aforementioned degradations and enhance the overall quality of the acquired images for

further image processing stage.

In the subsequent section, a comprehensive summary of the state-of-the-art image enhancement techniques dealing with noise, blur, and uneven illumination for WCE images is presented.

4.3.2 The SOTA image denoising methods

Noise is a prevalent distortion often encountered in video acquisition and transmission systems. Denoising, the process of reducing or removing noise, has been a subject of active research not only in the medical imaging domain but also in various fields of image processing for a substantial period. State-of-the-art denoising techniques can be categorized into three primary approaches: classical methods, dictionary learning-based methods, and deep learning-based methods.

4.3.2.1 Classical methods

Image restoration techniques can be categorized into three two groups: the filtering-based methods and the model-based methods. In this subsection, we make a comprehensive overview of the state-of-the-art of these groups.

Filtering-based image restoration To enhance the efficacy of denoising, more advanced kernels have been introduced. These encompass the bilateral filter [94], anisotropic filtering [95], [96], SUSAN filter [97], least-mean-square adaptive filter [98], trained filter [99], steering kernel regression [100], and edge-guided interpolation kernels [101]–[104], among others. These kernels are designed to exploit the local structure of images, including edge directions, in order to enhance the reconstruction of degraded images.

Filtering-based methods operate on the fundamental concept of modeling the relationships between neighboring pixels within an image using a local filter. The coefficients of this filter are determined based on the spatial distance between these pixels. To estimate the true value of a pixel in a degraded image, a weighted average of its neighboring pixels is calculated, with the weights being determined by the coefficients of the local filter kernel or a simple function. Over the past few decades, numerous researchers have developed various local kernels for image restoration. These include the Gaussian filter [105], Gabor filter [106], geometric filter [20], as well as fixed-function kernels like nearest, bilinear, and bicubic interpolation [107], [108] for image denoising.

While local filtering methods are known for their simplicity and ease of implemen-

tation, they come with a notable drawback. They are highly susceptible to severe degradation factors, particularly when dealing with high levels of noise or substantial upscaling needs. This susceptibility is a result of the significant disruption of the correlation between adjacent pixels, which hinders the accurate estimation of pixels in the latent image.

Model-based image restoration One distinctive feature of images is their total variation (TV), which quantifies the level of intensity variations among adjacent pixels. It is determined by summing the absolute values of the image gradient. In high-quality images, the TV tends to be minimal. However, when an image undergoes degradation, such as being affected by noise, its TV experiences a significant rise.

A specific category of representation techniques, known as TV-based methods, utilizes the total variation (TV) of images as a prior model (denoted as $p(\mathbf{x})$) to impose regularization on the TV of the degraded image (denoted as $\mathbf{y}$). The goal is to bring the TV of the reconstructed image in line with that of the reference image ($\mathbf{x}$) [109], [110]. These methods integrate the TV as a regularization component with the intention of reinstating the TV attributes of the original image.

Total variation (TV) constitutes a well-established and early-developed category of techniques employed in image enhancement. It measures the image's gradient, effectively capturing intensity variations between neighboring pixels. Mathematically, the total variation of an image $\mathbf{x}$ can be expressed as follows:

$$TV(\mathbf{x}) = \|\boldsymbol{\mathcal{D}}\mathbf{x}\|_1, \tag{2.25}$$

where $\boldsymbol{\mathcal{D}}$ denotes the gradient operator.

In the pioneering work on TV restoration by Rudin et al. [109], the TV was utilized as a prior model, denoted as $p(\mathbf{x})$, to regulate the denoising process. The objective was to minimize the TV of the reconstructed image, and it was formulated as follows:

$$\arg\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda \cdot \mathrm{TV}(\mathbf{x}) \right\}, \tag{2.26}$$

where the degraded image is denoted as $\mathbf{y}$, while $\mathbf{x}$ is used to represent the reconstructed image. The first component is the data fidelity term, which aims to align the reconstructed image with the degraded image. The second component involves the total variation (TV) of the reconstructed image, introducing regularization to promote smoothness and reduce noise. The parameter λ controls the trade-off between data fidelity and the desired level of smoothness.

Although TV methods have proven effective, they are not without their limitations. One significant drawback is the emergence of staircase artifacts. To tackle this challenge, several works [111]–[114] have introduced variants of the TV model. These adjustments involve the inclusion of higher-order image derivatives to mitigate staircase artifacts and yield more visually attractive reconstructed outcomes.

Another drawback of the TV model is its failure to account for the orientation of image gradients, rendering it less suitable for images with intricate textures exhibiting dominant directions in edges and contours. To surmount this limitation, Bayram and Kamasak [115] introduced the directional total-variation (DTV) method. DTV incorporates weights into the image gradient coefficients based on their respective directions, leading to a notable enhancement in denoising performance for images with natural textures characterized by prominent directions.

Instead of working on the examination of the prior model $p(\mathbf{x})$ within the image domain, many researchers decided to work on the analysis of wavelet coefficient distributions. This leads to the introduction of prior models within the wavelet domain. The core concept of the wavelet-based approach implies the representation of an image $\mathbf{x} \in \mathbb{R}^N$ as a linear combination of orthogonal basis functions. These functions are derived from a mother wavelet and involve different dilations and translations.

The wavelet transform has proven highly effective for image noise reduction tasks. When dealing with image denoising, if we assume that the noise model involves an identity matrix denoted as $\mathbf{H}$, we can express the noisy image as $\mathbf{y} = \mathbf{x} + \boldsymbol{\eta}$. Upon applying the wavelet transform to these image datasets, the following results emerge:

$$\mathbf{W}\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{W}\boldsymbol{\eta} = \boldsymbol{\alpha}_w + \mathbf{v}, \tag{2.27}$$

where $\mathbf{W}\mathbf{y} \in \mathbb{R}^N$ signifies the wavelet coefficients of the noisy image, while $\boldsymbol{\alpha}_w$ and $\mathbf{v}$ represent the wavelet coefficients corresponding to the unknown clean image $\mathbf{x}$ and the residual noise, respectively. The primary aim of wavelet-based denoising involves the estimation of $\hat{\boldsymbol{\alpha}}_w$ based on the wavelet transformation of the noisy image, denoted as $\mathbf{W}\mathbf{y}$. The recovered image $\hat{\mathbf{x}}$ can subsequently be reconstructed by employing the inverse transform, expressed as $\hat{\mathbf{x}} = \mathbf{W}^{-1}\hat{\boldsymbol{\alpha}}_w$.

In wavelet-based denoising algorithms, one often encounters the utilization of shrinkage techniques, including soft or hard thresholding. These approaches serve to diminish or eliminate wavelet coefficients falling below a specified threshold. While this thresholding effectively mitigates noise, it can also lead to the deletion or substantial reduction of minor wavelet coefficients that correspond to intricate image details. Consequently,

the reconstructed image may fail to preserve the intended degree of detail and instead display an excessively smoothed appearance.

4.3.2.2 Dictionary learning-based methods

Another strategy involves representing images in an alternative linear space to efficiently capture their probabilistic characteristics within image structures. Typically, a transformation domain is designed based on specific assumptions, ensuring that the image or signal representation adheres to particular probabilistic criteria. For instance, total-variation methods assume images are piecewise smooth with sparsely distributed gradients. Meanwhile, the wavelet transform is formed through image representation in an orthogonal wavelet dictionary, where wavelet coefficients feature sharp peaks centered around zero and follow a heavy-tailed distribution. Various works in the literature have considered mathematical models like the Laplacian, generalized Gaussian, Gaussian scale mixture, etc., to emulate the empirical distribution of images. However, addressing the computational challenges associated with learning priors and conducting optimization across the entire image domain remains a significant concern. Furthermore, these approaches frequently result in excessive smoothing and the blurring of details in the reconstructed image. This implies that an image can often be characterized using a limited selection of basis functions chosen from a larger set. Investigations in certain domains, such as wavelets and gradients, have provided evidence of the sparse representation of images. This observation has served as an inspiration for numerous research endeavors focused on image reconstruction.

In these methodologies, an image can be regarded as either an isolated signal or in relation to other corresponding images within a dictionary. This flexibility permits reconstruction to be executed either independently for each image or collaboratively with neighboring images. Let's designate the dictionary as $\{\mathbf{y}_i|\mathbf{y}_i \in \mathbb{R}^N; i = 1,\ldots,N\}$, encompassing N images. The primary aim is to determine a latent representation $\mathbf{x}_i \in \mathbb{R}^n$ for each image $\mathbf{y}_i$, with the condition:

$$\mathbf{y}_i = \mathbf{H}_i\mathbf{x}_i + \boldsymbol{\eta}_i, \tag{2.28}$$

where $\boldsymbol{\eta}_i$ is the additive noise in the image $\mathbf{y}_i$, and $\mathbf{H}_i$ denotes the degradation matrix on $\mathbf{x}_i$.

Without loss of generality, the underlying motivation of these approaches is to express each latent representation $\mathbf{x}_i$ as a linear combination of K basis vectors, often referred to as "atoms," denoted as $\{\mathbf{d}_1,\ldots,\mathbf{d}_j\ldots,\mathbf{d}_K|\mathbf{d}_j \in \mathbb{R}^n\}$ from a dictionary of images $\mathbf{D} \in \mathbb{R}^{n\times K}$. This can be expressed as $\mathbf{x}_i = \mathbf{D}\boldsymbol{\alpha}_i$, where $\boldsymbol{\alpha}_i \in \mathbb{R}^K$ represents the

vector of representation coefficients for $\mathbf{x}_i$. The problem presented in (2.28) can be reformulated as:

$$\mathbf{y}_i = \mathbf{H}_i\mathbf{D}\boldsymbol{\alpha}_i + \boldsymbol{\eta}_i \tag{2.29}$$

The restoration of a degraded image $\mathbf{y}_i$ is equivalent to estimating a coefficients vector $\boldsymbol{\alpha}_i$ that satisfies the degradation model. Under the MAP framework and Gaussian noise assumption, we have:

$$\hat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{\alpha}_i}\{||\mathbf{y}_i - \mathbf{H}_i\mathbf{D}\boldsymbol{\alpha}_i||_2^2 - \lambda\log\left(p(\boldsymbol{\alpha}_i)\right)\} = \arg\min_{\boldsymbol{\alpha}_i}\{||\mathbf{y}_i - \mathbf{H}_i\mathbf{D}\boldsymbol{\alpha}_i||_2^2 + \lambda\Phi(\boldsymbol{\alpha}_i)\} \tag{2.30}$$

In the literature, substantial efforts have been dedicated to the development of adaptive dictionaries suited to data-specific attributes. Researchers have explored various methods, including Principal Component Analysis (PCA)[116]–[119], sparse learning [120]–[126], and combinations of PCA and sparsity [127]–[129]. These techniques are aimed at building dictionaries precisely customized to match the traits inherent in the provided dataset. Nevertheless, despite their commendable performance spanning a range of enhancement applications, a limitation of dictionary models lies in their potential inability to accurately represent the genuine image distribution within the vector space constructed by the dictionary atoms $\mathbf{D}$.

To overcome this constraint, it becomes crucial to investigate alternative prior models capable of more accurately describing the true image distribution. By designing a prior model that better fits to the underlying distribution, the potential for enhancing the quality of the reconstructed image is heightened.

4.3.2.3 Deep-learning-based methods

Over the past few years, the field of image restoration has witnessed a significant rise in the popularity of deep learning methods, due to the significant development of artificial intelligence and computer vision. These methodologies seek to acquire a compact inference or mapping function by training on collections of paired images, where one is degraded, and the other is latent. This function is then applied to recover the original image. The essential concept behind this learning process can be expressed as follows:

$$\arg\min_{\boldsymbol{\Theta}} \mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}) \quad \text{s.t.} \quad \hat{\mathbf{x}} = \mathcal{F}(\mathbf{y}, \mathbf{H}; \boldsymbol{\Theta}), \tag{2.31}$$

where $\mathcal{F}(\cdot)$ symbolizes the inference or mapping function, with its parameters defined by $\boldsymbol{\Theta}$, while $\mathcal{L}(\cdot)$ denotes the loss function utilized to evaluate the similarity between the reconstructed image $\hat{\mathbf{x}}$ and the reference image $\mathbf{x}$ from the training dataset. Deep neural networks, including convolutional neural networks (CNN)[130]–[138], multi-layer

perceptrons[139], and stacked sparse denoising autoencoders [140], [141], are often employed to represent the mapping function $\mathcal{F}(\cdot)$.

Once the training phase is completed, and the parameter set $\mathbf{\Theta}$ of the neural network is obtained, the previously unknown latent image $\hat{\mathbf{x}}$ can be reconstructed from its degraded observation $\mathbf{y}$ using the expression $\hat{\mathbf{x}} = \mathcal{F}(\mathbf{y}, \mathbf{H}; \mathbf{\Theta})$. In the past decade, Convolutional Neural Networks (CNN) have gained popularity for deep noise removal[142], [143], primarily due to their capability to address complex noise scenarios. However, they rely heavily on having high-quality noisy-clean image pairs for training, which is a challenging endeavor. To better retain fine spatial details, CNN-based multiscale techniques[144], [145] have been proposed. Despite the remarkable performance of deep learning in image restoration, it does have a limitation. This limitation comes from the fact that the parameters $\mathbf{\Theta}$ and the mapping function $\mathcal{F}(\ldots)$ are tailored to specific degradation processes, such as particular noise levels. Consequently, this limits the adaptability of the trained models for diverse image enhancement tasks.

The deep learning methodology entails the training of a model using a designated dataset comprising images affected by a particular form and level of degradation. The model acquires the ability to associate deteriorated observations with their corresponding noise-free or latent counterparts. Nevertheless, when dealing with an alternative kind or extent of degradation, the trained model might not deliver optimal results or may not be suitable at all. This discrepancy arises from the fact that the model's parameters and mapping functions are intricately tailored to fit the attributes of the training dataset, encompassing the precise degradation process employed.

As a result, using the trained models for distinct image enhancement tasks involving different forms or distortion severity levels presents a notable challenge. In such cases, there is a need to either retrain the model using fresh data that reflects the specific degradation in question or create more versatile models capable of adeptly addressing a wider spectrum of degradation scenarios. Mitigating the constraint associated with deep-model learning's adaptability to various degradation processes continues to be a vibrant area of research, enabling them to effectively manage a multitude of restoration tasks.

In the context of WCE image denoising, some of the most famous traditional image processing methods applied to WCE images encompass the geometric mean filter[20], Total Variation (TV)[110], and the wavelet-based approach[146]. More recently, a novel approach[19] has emerged, which involves the utilization of the Deep Image Prior (DIP) technique, followed by the deployment of a blind image quality assessment network. This

approach iteratively generates a noisy image from a given noise model and subsequently produces a clear image from the noisy one. However, a significant challenge to be noted is that this process requires a substantial number of iterations to achieve an optimal reconstructed image from the noise model. Consequently, the described procedure is time-intensive and lacks efficiency in practical applications.

4.3.3 The SOTA image deblurring methods

Similar to denoising, much study has been done in recent decades with many different image deblurring solutions on both natural and medical imaging domains. Regarding natural images, image deblurring methods can be divided into classical image processing-based methods and CNN-based ones.

4.3.3.1 Classical methods

Historically, traditional image deblurring techniques have included Wiener filtering [147], the Lucy-Richardson algorithm [148], and the total variation algorithm [149]. These algorithms primarily address non-blind image deblurring and often fail to fully leverage the inherent prior information found in natural images. Consequently, the resulting restored images may exhibit a lack of precision.

Recognizing the limitations inherent in classic deblurring techniques, subsequent traditional methods have endeavored to exploit the prior information found in natural images to enhance deblurring performance. These approaches typically extend upon the foundation of classic algorithms by incorporating additional priors, such as the sparsity commonly observed in natural images. By making use of the statistical attributes inherent to natural images, these priors are employed as regularization terms to augment the quality of deblurring results.

Furthermore, traditional methods also involve the utilization of blind deconvolution techniques, which concurrently estimate both the original image and the blur kernel. In 2006, Fergus et al. [150] introduced a pioneering non-parametric motion blur kernel estimation technique, marking an early instance of blind image deblurring. Their method encompassed the learning of a mixed Gaussian image prior and a mixed exponential fuzzy kernel prior. Leveraging a variational bayesian (VB) criterion, the authors introduced a non-parametric image deconvolution approach. The objective of this approach was to estimate both the underlying image and the blur kernel without relying on prior knowledge of specific blur characteristics.

4.3.3.2 Deep-learning-based methods

In the domain of image deblurring, CNN-based methodologies, including Generative Adversarial Networks (GANs), have also found practical application. GANs have been deployed to grasp the end-to-end regression mappings, effectively connecting a blurred input image with its corresponding high-quality image [151], [152]. However, it's worth noting that in the medical domain, employing generative models carries the potential risk of introducing uncontrolled elements that could lead to inaccurate diagnostic outcomes.

To tackle this challenge, Zhang et al. [153] introduced an innovative hierarchical representation strategy that utilizes a multi-patch network inspired by spatial pyramid matching. This approach adeptly addresses blurred images across various scales by integrating a network structure that accounts for multiple patches and their corresponding spatial relationships.

In a more recent development, Wang et al. [154] integrated deep U-net networks with a self-attention mechanism. This innovation allows for the efficient extraction of local contextual information from degraded images, empowering the network to more effectively manage the deblurring task.

These progressions in CNN-based approaches underscore the ongoing efforts to enhance image deblurring methodologies. Researchers persistently explore novel network architectures, incorporate attention mechanisms, and take into account the specific demands of medical imaging to enhance the precision and efficiency of deblurring algorithms in the medical domain.

Within the domain of WCE image deblurring, researchers have dedicated their efforts to create specialized techniques customized for this particular field. Liu et al. [155] introduced a modification to the standard parameter updating iteration in total variation, incorporating the fast iterative shrinkage/thresholding technique. This adaptation aimed to enhance the capability to manage multichannel WCE images and elevate the overall deblurring effectiveness.

In recent investigations, researchers have turned to dictionary learning approaches to tackle the challenges associated with WCE image deblurring. Wang et al. [156] applied dictionary learning to extract the inter-frame relationships within WCE videos, primarily for super-resolution applications. By exploiting the inherent redundancies and connections among consecutive frames, they succeeded in enhancing the resolution of the reconstructed images. Similarly, Peng et al. [157] used the dictionary learning techniques for generating deblurred reconstructions in WCE images. This approach

aimed to capture the underlying structure and patterns within the image content, thus advancing the deblurring process.

These developments highlight the utilization of adapted techniques, including adapted total variation and dictionary learning, to deal with the distinctive features and difficulties inherent in WCE image deblurring. These approaches consider the particular prerequisites and limitations of WCE imaging systems, thereby enhancing the overall quality of the reconstructed images.

4.3.4 The SOTA uneven illumination correction methods

Uneven illumination is a common problem encountered in digital images captured by sensors. This issue arises due to the flat nature of digital sensors used in cameras. Pixels situated at the sensor's center receive light rays directly, coming in at a 90-degree perpendicular angle. However, pixels located at the corners of the sensor receive light rays at a slight angle. This difference in the angle at which light enters results in a gradual decrease in pixel intensity as you move from the image's center to its edges. Unevenly illuminated images are obtained using the following formula:

$$\mathbf{y}_i = \mathbf{x}_i + \mathbf{H}, \tag{2.32}$$

where the unevenly illuminated image $\mathbf{y}_i$ is derived from the pristine source image $\mathbf{x}_i$ under ideal conditions. Meanwhile, the background image $\mathbf{H}$ is extracted through a series of steps, starting with a low-pass filter and subsequent subtraction from the source image. This subtraction process aims to harmonize the contrast and mitigate the impacts of uneven illumination. To further compensate for these effects, the subtracted image undergoes a contrast stretching procedure. Similarly, two prevalent methods, classical and deep learning-based correction, are employed to rectify the uneven illumination. Classical techniques assume the presence of a uniform background in the image, occasionally containing smaller, brighter, or darker objects.

4.3.4.1 Classical methods

In the classical method, the prospective correction for uneven illumination involves integrating two specific types of additional images taken during the image capture process. The first type encompasses dark images, capturing the background in the absence of any light, while the second type includes bright images taken with the object removed from the background. Multiple images from these two sets are acquired to minimize noise and address any lighting irregularities [158], [159]. Following this, a

transformation function is applied to generate a corrected image.

$$\mathbf{x}_i = \frac{\mathbf{o}_i - \mathbf{d}_i}{\mathbf{b}_i - \mathbf{d}_i} * C, \tag{2.33}$$

where $\mathbf{o}_i$, $i = 1 \ldots N$, is the source image, N is the number of images, $\mathbf{d}_i$ and $\mathbf{b}_i$ represent the dark image and bright image, respectively. C is a normalization constant used for the purpose of recovering the source image color [160], which is computed as:

$$C = \frac{Mean(\mathbf{o}_i)}{Mean\left(\frac{\mathbf{o}_i - \mathbf{d}_i}{\mathbf{b}_i - \mathbf{d}_i}\right)} \tag{2.34}$$

Numerous investigations have delved into perspective correction techniques for document imaging. In [161], an approach is detailed for rectifying perspective distortion in document images, with a particular emphasis on uncalibrated images. [162] presents a method that relies on fuzzy sets and morphological operations to eliminate perspective distortion and restore a front-parallel view of text using a single image. In another work [163], the focal point is on distorting information and rectifying perspective in Chinese document images without any prior knowledge of the image's edge or paragraph format. This particular study makes use of the vertical strokes and horizontal characteristics of Chinese characters to facilitate image correction. Additionally, [164] addresses skew correction and perspective rectification in camera-based Farsi document images.

Within the domain of medical image processing, the correction of uneven illumination has been a persistent challenge. Over time, several classical techniques have emerged to tackle this problem. These approaches encompass filtering methods [165], [166] as well as models based on the Retinex principle [167].

Filtering-based approaches, as demonstrated by the work of Leong et al. [165] and Wang et al. [166], employ diverse filtering operations to alleviate the influence of uneven illumination in medical images. The primary objective of these methods is to improve the overall image quality and enhance the discernibility of structures by reducing the negative effects of uneven illumination.

Retinex-based models, such as the one proposed by Han et al. [167], are inspired by the retinex theory, which seeks to replicate the HVS's ability to perceive the relative reflectance of objects under varying lighting conditions. These models aim to estimate and rectify the uneven illumination present in medical images, resulting in an improved image quality and more accurate diagnoses.

These traditional methods play a pivotal role in addressing the challenges posed by uneven illumination in medical image processing. However, it's worth noting that

the field is continuously evolving, and researchers are actively exploring CNN-based techniques and approaches to further enhance uneven illumination correction in medical images.

4.3.4.2 Deep-learning-based methods

Several deep learning-based approaches have been proposed to tackle the challenge of uneven illumination correction in both natural and medical image processing. Deep learning techniques have exhibited impressive capabilities in grasping intricate image patterns and extracting meaningful features, rendering them particularly well-suited for handling demanding tasks such as uneven illumination correction. These methods exploit the potency of deep neural networks to effectively learn the relationship between the input image and its corresponding corrected version.

For instance, [168] introduces an innovative approach to deal with non-uniform illumination in underwater images using a fully convolutional network (FCN). This method aims to enhance the visual quality of underwater images by proficiently eliminating the undesirable effects of varying illumination, which often lead to diminished contrast and visibility. By exploiting the potential of FCN, this approach efficiently learns to rectify the non-uniform illumination and restore the genuine colors and details in underwater scenes. In another work, [169] tackles the issue of nonuniform illumination in underwater images through a maximum likelihood estimation (MLE) strategy. Additionally, [170] addresses this concern by scrutinizing the color distribution across the mosaic and devising a correction model grounded in statistical analysis.

Despite significant efforts, there are few methods available to tackle challenges in WCE images. Long et al. [171] introduced a technique that relies on high-quality guided images of the same scene to enhance the quality of low-quality WCE images through histogram comparison. Their adaptive fraction-power transformation approach dynamically adjusts the power value based on local image characteristics to achieve the best possible enhancement outcomes. However, this method may face constraints when guided images are absent or when there is substantial dissimilarity between the guided and low-quality WCE images. Such differences could potentially introduce unwanted artifacts and compromise diagnostic accuracy.

To address this concern, Prasath et al. [16] introduced a customized variant of the feature-linking model (FLM) [172]. This adapted method seeks to enhance image quality by converting RGB images into the HSV color space, leading to improved chromaticity and the avoidance of streaking artifacts often associated with histogram-based models.

Nevertheless, it's worth noting that FLM relies on linking image features across multiple scales, which may not be optimal for enhancing images with low contrast or intricate backgrounds. Additionally, the computational complexity inherent in FLM could pose challenges for real-time applications that require rapid image processing.

It's worth noting that low-light enhancement and uneven illumination are separate problems that can significantly degrade image quality. They require distinct correction techniques for enhancement. In recent times, several low-light enhancement approaches have been proposed for natural images, such as LIME [173], EnlightenGAN [174], and Deep Retinex [175]. In this study, our objective is to evaluate the performance of these methods in addressing the specific problem of uneven illumination in wireless capsule endoscopy images.

4.3.5 Discussion

While the methods discussed earlier offer solutions for common issues like low light conditions, noise, or low contrast, they do come with their own set of difficulties. To elaborate, gamma correction techniques may lead to an undesirable over-amplification of pixel values, potentially introducing various unwanted artifacts into the processed images. Furthermore, in cases where an image exhibits a wide dynamic range or contains regions with extremely high or low pixel values, histogram equalization can result in the creation of non-photorealistic areas and generate images that appear artificial or unnatural.

Considering the challenges previously mentioned, recent advancements in deep learning, particularly within the domain of deep convolutional neural networks (CNNs), have laid the groundwork for the development of novel image enhancement solutions. Modern deep learning models, trained on large-scale medical datasets to acquire generalizable insights, typically fall into one of two architectural categories: encoder-decoder or generative models. Encoder-decoder models systematically map input data to a latent embedding space before gradually restoring it to the original resolution. While these methods effectively reduce input dimensionality to capture essential features, they tend to discard global contextual information, making image reconstruction challenging. Conversely, generative networks process images with more detailed spatial features but may generate uncontrolled features less suitable for precise medical tasks such as pathological classification. Recently, attention-based methods have proven effective in a variety of applications by leveraging positional information and inter-pixel relationships to address these limitations.

4.4 The SOTA WCE image classification methods

4.4.1 Introduction

In this section, we aim to present the state-of-the-art classification techniques for WCE images, focusing on two main aspects: mono-pathology and multi-pathology classification. While mono-pathology classification refers to the task of categorizing WCE images as to whether they belong to one specific class corresponding to single diseases or abnormalities (binary classification, e.g. bleeding, polyps, tumor, or ulcer), multi-pathology classification involves the simultaneous identification of whether it belongs to one in a list of trained pathologies. This task presents additional challenges due to the complexity of detecting and differentiating multiple pathologies accurately.

4.4.2 Mono-pathology classification

Mono-pathology classification involves the task of classifying WCE images into specific categories that correspond to individual diseases or abnormalities, typically adopting a binary classification approach. The primary objective is to accurately ascertain whether an image relates to a particular pathology or not. In recent times, substantial progress has been achieved in mono-pathology classification by making use of diverse machine learning techniques, particularly deep learning.

In [176], Maghsoudi et al. employed a method that utilized color, geometry (invariant moments), and texture features (GLCM, Gabor, LBP, and Laws' features). Unlike LBP, they selected the local frequency pattern (LFP) technique to extract features, considering the concept of partial truth. In the case of Iakovidis et al. [177], features were extracted through a CNN, and the deep saliency detection (DSD) algorithm was utilized to identify significant points associated with GI anomalies in frames from endoscopic videos. Guo et al. [178] fine-tuned EfficientNet for feature extraction from endoscopic images and fused spatial and channel features. They subsequently employed an attention network as a classifier for binary classification. Noya et al. [179] employed a fusion of image processing and machine learning techniques to extract features from wireless capsule endoscopy images, encompassing both color and texture information. These features were then fed into a boosted decision tree classifier. In the research by Leenhardt et al. [180], local features were derived through the training of a neural network on a dataset containing images with known pathology findings. These locally extracted features captured the unique characteristics and patterns specific to angiectasia lesions. The algorithm then employed these extracted local features within a classification network.

Firstly, concerning bleeding detection, color and texture features are also extracted, as explained in the study by Usman et al. [181]. Following this, the region of interest (ROI) can be chosen either manually, as demonstrated in the work of Pan et al. [182], or through the use of a region-growing algorithm, as illustrated in Sainju et al.'s approach [183]. Furthermore, various transformations are considered to enhance the contrast between the bleeding lesion and the surrounding normal mucosa region, as exemplified in the research by Ghosh et al. [184]. For instance, in the studies conducted by Figueiredo et al. [185] and Kundu et al. [186], a smoothing enhancement is carried out on the second component of the CIE Lab transformed WCE image to amplify the contrast between the bleeding lesion and the surrounding normal mucosa region. Subsequently, these improved images are input into classifiers like Support Vector Machines (SVM), as demonstrated in Ghosh et al. [187] and Fu et al. [188], or unsupervised-learning clustering, as discussed in Jia et al. [189] and Bchir et al. [190]. Additionally, various color spaces or coordinate systems are employed. For instance, in Xing et al.'s work [191], the image is transformed into the polar coordinate system, and a CNN is used to generate the saliency map. This saliency map is then combined with generated features for classification. In Hajabdollahi et al.'s work [192], the image is converted into different color spaces, and a Multilayer Perceptron (MLP) is applied to the concatenated channels.

Secondly, to detect polyps, in the study referenced as [193], a combination of Gabor texture features is employed alongside K-means clustering, followed by the extraction of geometric features. Similarly, in [194], Gabor texture features are paired with the SUSAN algorithm. Yuan and Meng [195] make use of the coordinates obtained from the shaded regions of Gabor filters and the Monogenic-Local Binary Pattern (M-LBP) to simplify component measurements. They utilize Linear Discriminant Analysis (LDA) for classification, along with support vector machines (SVM). Conversely, other researchers opt for invariant mapping algorithms such as the Scale-Invariant Feature Transform (SIFT) or Histogram of Oriented Gradients (HOG) for feature extraction, as observed in studies like [196]–[198]. These extracted features are subsequently input into classifiers, notably SVM, as illustrated in [199]–[201].

In recent times, methods based on CNNs have gained prominence as effective strategies for feature extraction and classification [202]. Tajbakhsh et al. [203] utilize a deep CNN that combines both the global geometric properties of polyps and the local intensity variation patterns along polyp boundaries to perform classification. Several studies, including [204]–[206], employ transfer learning techniques on the WCE dataset. These methods leverage pre-trained networks or utilize inter-frame information to

estimate the polyp region. In such cases, a tracker that has been pre-trained on the YOLO Residual network is often used. In [207], Yuan proposes an unsupervised deep CNN model that uses stacked autoencoders to learn features from the latent space and subsequently classify polyp images in WCE images.

Thirdly, in the tumor detection field, various methodologies have been applied. Wavelet transform has been used in research studies such as [208]–[210], where it is employed to extract texture features from the images. Similarly, the discrete cosine transform (DCT) is utilized for texture feature extraction in [211], while the curvelet transform is exploited in [212]. Following this, support vector machines (SVM) are commonly employed for classification. In the study conducted by Vieira et al. [213], they employ a multivariate Gaussian model to estimate the different distributions between tumor and normal pixels. The maximum a posteriori probability (MAP) is subsequently used to evaluate the likelihood between pixels in the CIE Lab color space channels, which are assumed to follow Gaussian distributions. In the work of Mahdi et al. [214], texture features like contrast energy are extracted and fed into an artificial neural network (ANN)-based Fuzzy Network for classification.

Finally, regarding ulcer detection, similar to the context of tumor identification, the extraction of color and texture features is frequently carried out using the Gabor filter and the discrete cosine transform (DCT) [215], [216]. Souaidi et al. [217] introduce a multi-scale method known as MS-CLBP (inclusive local binary pattern and Laplacian pyramid) to classify lesions from a multi-scale perspective. In another study by Wang et al. [218], RetinaNet is employed for the initial detection of ulcers. Following this, based on the size of the identified region, the segmented patch is input into two separate networks to further validate the accuracy of the detection. Moreover, some pre-trained networks such as X-ception CNN [219], [220] are utilized for feature extraction, followed by the utilization of Random Forest for classification purposes.

In the subsequent subsection, we will provide the bibliography of the multi-pathology classification.

4.4.3 Multi-pathology classification

Accurately categorizing medical images is a vital aspect of pathology diagnosis, with the classification of multiple pathologies within a single image being particularly challenging. The presence of diverse pathologies increases the classification difficulty, demanding advanced techniques for thorough analysis and interpretation. In this section, we concentrate on the issue of multi-pathology classification and investigate the various

strategies and methods utilized in the literature to tackle this demanding task.

In the study conducted by Sekuboyina et al. [221], the initial step involves the conversion of endoscopy images into the CIE-Lab and YCbCr color spaces. Subsequently, the a-channel of the CIE-Lab space is employed as the input for a convolutional neural network (CNN) to perform classification. To tackle the issue of imbalanced data, the Synthetic Minority Over-sampling Technique (SMOTE) is applied as a method for oversampling the minority class. Sadasivan et al. [222] approach the task of multi-pathology classification by partitioning wireless capsule endoscopy (WCE) images into multiple patches. Each patch is identified using color and texture features extracted from the chromatic components within the CIE Lab color space. Yuan et al. [223] adopt a similar strategy, focusing on the extraction of color features from WCE images. They employ K-means clustering to categorize the extracted features into different groups and utilize dictionary learning for subsequent classification. In a related work by Nawarathna et al. [224], texture features are extracted through a combination of the Leung and Malik (LM) filter bank and a set of local binary patterns. The classification task is then addressed using the K-nearest Neighbor algorithm.

Deep learning-based methodologies have demonstrated their effectiveness in the realm of multi-pathology classification. Nadeem et al. [225] opt to extract a mixture of texture features, encompassing attributes like LBP and Haralick features, in conjunction with conventional characteristics like joint composite descriptor, auto color correlogram, color layout, and edge histogram. This combination of diverse features is then consolidated and funneled into a CNN for the classification task. In another approach, Lan et al. [226] introduced a profound cascade network referred to as "Cascade-Proposal" for multi-pathology classification. Their methodology entails the generation of a limited count of region proposals characterized by high recall, which is achieved through a module designed for rejecting region proposals. Concurrently, it identifies abnormal patterns using a specialized detection module. To further enhance precision, they exploit the multi-regional combination (MRC) technique to acquire regions of interest and apply the salient region segmentation (SRS) method for the precise localization of significant regions. Furthermore, they incorporate the dense region fusion (DRF) method to refine the boundaries of objects.

In recent years, attention methods have surfaced as a leading strategy for enhancing the classification of multiple pathologies. These attention mechanisms empower models to choose specific areas or attributes within an image, ultimately refining the classification performance by emphasizing pertinent details while suppressing irrelevant or noisy signals.

Several contemporary investigations have delved into applying attention mechanisms within this context. Xing et al. [227], for instance, introduced an approach that combines attention mechanisms with deformation to magnify areas of interest within lesions. They derive attention maps through self-attention and subsequently employ deformation techniques to magnify the pinpointed lesion region. By contrasting the self-attention maps before and after deformation, they produce a conclusive attention map for the purpose of classification.

To tackle the issue of bias concerning value in attention loss, Guo and Yuan [228] introduced the utilization of angular loss in the realm of classification. They implemented attention mechanisms to produce diverse attention maps based on multi-level characteristics. Furthermore, they integrated a semi-supervised method to manage data lacking labels. Similarly, Zhao et al. [229] used multiple attention layers and adopted the adaptive cosine loss to refine the classification process. Their objective in incorporating attention mechanisms was to amplify the model's capacity for discrimination.

These studies underscore the efficacy of attention methodologies in multi-pathology classification. Attention mechanisms empower the model to concentrate selectively on informative regions or characteristics, thereby augmenting classification precision and resilience. Furthermore, these approaches tackle distinct challenges like value bias and untagged data through inventive techniques like angular loss and semi-supervised learning.

4.4.4 Discussion

In recent years, attention techniques have emerged as a prominent approach for improving multi-pathology classification tasks. Attention mechanisms enable models to selectively focus on important regions or features within an image, enhancing overall classification performance by highlighting relevant information while suppressing irrelevant or noisy signals. These studies highlight the effectiveness of attention techniques in multi-pathology classification. Attention mechanisms allow for selective focus on informative regions or features, thereby improving classification accuracy and robustness. Additionally, to address specific challenges such as value bias and unlabeled data, innovative approaches such as self-supervised learning could be considered.

5 Conclusion

In this chapter, we briefly presented a survey of the state-of-the-art approaches including the IQA methods, the quality enhancement methods, and the image classification

methods. To enhance our comprehension and facilitate appropriate quality enhancement, we will commence by elucidating different prevalent distortion models. This will furnish us with an in-depth understanding of the distortions and pave the way for more effective quality improvement strategies.

Among this state-of-the-art study, we also observed that the attention-based approach, while it has been well developed for both image enhancement and image classification with very promising results, is not satisfactorily investigated. This would be interesting to develop learning-based approaches and see to what extent these approaches can be applied to different types of tasks in WCE image processing. It's important to highlight that recent progress in image restoration and enhancement techniques, particularly those leveraging machine learning, relies heavily on access to datasets containing both corrupted and clean image pairs for effective training. In the context of WCE, however, a noteworthy challenge persists due to the absence of a dedicated quality assessment dataset. To address this challenge, it becomes imperative to introduce a quality assessment dataset featuring WCE videos exhibiting various degrees of distortions. Such a dataset would play a pivotal role in advancing the development of precise and dependable image enhancement algorithms, ultimately enhancing the diagnostic potential of WCE systems. For this purpose, in the next chapter, we will first introduce the proposed distortion dataset for WCE images.

CHAPTER 3

The proposed distortion dataset: *A Quality-Oriented Database for Video Capsule Endoscopy (QVCED)*

Abstract

In this work, we propose a novel dataset called the Quality-Oriented Database for Video Capsule Endoscopy (QVCED) which serves as the primary crucial resource for evaluating the quality of Wireless Capsule Endoscopy (WCE) images and videos. Serving as a benchmark, the QVCED encourages the design of learning-based enhancement methods to address image quality assessment and enhancement challenges in WCE. This comprehensive dataset consists of a large number of WCE videos encompassing common distortions encountered in clinical practice, including noise, defocus blur, motion blur, and uneven illumination. Moreover, video quality has been intentionally degraded at varying distortion severity levels to faithfully replicate real-world conditions. The extensive analysis demonstrates the diversity and practical relevance of this dataset in the WCE domain that motivates the advancement of a more precise diagnosis regarding gastrointestinal disorders. The complete dataset is publicly available through the following link: https://cloud.math.univ-paris13.fr/index.php/s/b74TQk7mMpHDXKT.[1]

[1]T. -S. Nguyen, J. Chaussard, M. Luong, A. Beghdadi, H. Zaag, and T. Le-Tien, *"A Quality-Oriented Database for Video Capsule Endoscopy,"* 11th European Workshop on Visual Information Processing (EUVIP), Norway, 2023 (Winner of Three-Minute Thesis (3MT) Competition Award).

Chapter content

1 Introduction 56

2 Proposed Dataset - QVCED 57

2.1 Reference Videos Selection 58

2.1.1 Noise Assessment 59

2.1.2 Blur Assessment 59

2.1.3 Uneven Illumination Assessment 60

2.2 Distortion Generation 60

2.2.1 Noise Model 60

2.2.2 Defocus Blur Model 62

2.2.3 Motion Blur Model 62

2.2.4 Uneven Illumination Model 64

3 Dataset Analysis 67

3.1 Subjective Test 69

3.1.1 Testing Environment 69

3.1.2 Video Quality Score 70

3.2 Diversity Data Analysis 72

4 Conclusion 73

1 Introduction

Wireless Capsule Endoscopy (WCE) has revolutionized medical practices for gastrointestinal (GI) disease screening and diagnosis [230]. However, a major challenge in WCE is obtaining optimal image quality, which directly affects diagnostic accuracy. Indeed, WCE image quality suffers from distortions due to the limitations of the sensor technology and the constrained acquisition environment. For example, narrow apertures and small sensors with limited dynamic range and sensitivity generate noise within captured frames [19]. Especially, additive white Gaussian noise in WCE images is the accepted standard model [20]. Unstable environments result in excessive blurriness [231] due to the uncontrolled and random motion of the capsule, while the capsule's limited lighting coverage causes uneven illumination[18]. These distortions can decrease the performance of tasks like lesion detection, recognition, and tracking in the gastrointestinal tract.

To address image quality limitation issues, due to the aforementioned distortions, numerous learning-based algorithms [16], [19], [156] have been proposed. Specifically, recent advancements in image restoration and enhancement techniques rely on learning-based methods that require pairs of corrupted and clean images for training. However, in the case of WCE, the absence of a dedicated quality assessment dataset poses a significant challenge. Therefore, a dataset specifically designed for assessing the quality of WCE images, with varying levels of distortions, is crucial for developing accurate and reliable image enhancement algorithms. To the best of our knowledge, there is currently no specialized dataset available specifically for assessing video quality in WCE. Existing datasets commonly used for quality assessment, such as LIVE Mobile VQA [232], KoNViD-1k[233], TID2013[234], CSIQ[235], and CID:IQ[236], have primarily focused on natural images for over two decades. In the field of medical imaging, datasets like RIQA[237] and LVQ[238] have been developed specifically for retinal and laparoscopic image/video evaluation, respectively. However, it is important to note that these datasets are not efficient for training learning-based quality enhancement techniques for WCE images due to the inherent dissimilarities in medical imaging types and modalities. Moreover, most medical databases are tailored for segmentation and classification tasks, making this work a valuable contribution to fulfilling a real requirement in medical imaging and in particular to evaluating and improving WCE image quality.

Consequently, toward the demand for a comprehensive video quality assessment dataset, we propose the Quality-Oriented Database for Video Capsule Endoscopy (QVCED), derived from the Kvasir-Capsule dataset [10]. QVCED covers a wide range of scenarios with different pathologies and multiple types of distortions, prioritizing

realistic conditions. The dataset is produced through a two-stage process. In the first stage, reference videos that meet the required quality criteria are carefully selected from the Kvasir-Capsule dataset [10]. Next, the reference video is subjected to a degradation process in which a controlled level of degradation is applied by means of the physical parameters of the used distortion generation model.

The subsequent sections of this chapter are structured as follows: Section 2 describes the creation process of the QVCED dataset. Within this section, Section 2.1 outlines the initial selection process for reference videos, while Section 2.2 explains the generation of simulated distortions in the chosen reference videos. Afterward, Section 3 focuses on the analysis and discussion of the proposed dataset. Specifically, Section 3.1 presents the implementation and results of a subjective test, including opinion scores from expert and non-expert observers regarding the simulated distortions. Furthermore, Section 3.2 analyzes the content diversity of the QVCED dataset. Finally, this chapter concludes in Section 4, summarizing the key findings and implications of this work in terms of the creation and analysis of the proposed dataset.

2 Proposed Dataset - QVCED

In this section, we describe the dataset creation process. We provide a comprehensive overview of the methodologies encompassing the selection of reference videos (Section 2.1) and the simulation of distortions applied to the chosen reference videos (Section 2.2).

In order to guarantee the strength and thoroughness of our dataset, we engaged in a careful process of selecting reference videos. In Section 2.1, we detail the criteria and methods we used to carefully choose the most representative and varied reference videos. This essential step in the dataset creation process was guided by a precise evaluation of video content, and quality.

After the thorough selection of reference videos, the second step is the distortion simulation, which plays a pivotal role in our dataset creation process. Section 2.2 focuses on explaining the details of our method for intentionally adding distortions to these reference videos. Our methodology involved a diverse range of distortion models, each thoughtfully designed to replicate real-world conditions and challenges.

More precisely, the first step is to select reference videos from an existing WCE dataset. We have created a dataset comprising 20 original reference videos extracted from the Kvasir-Capsule dataset[10]. These videos have a duration of 10 seconds, a resolution of 336×336 pixels, and a frame rate of 30 frames per second (fps). The

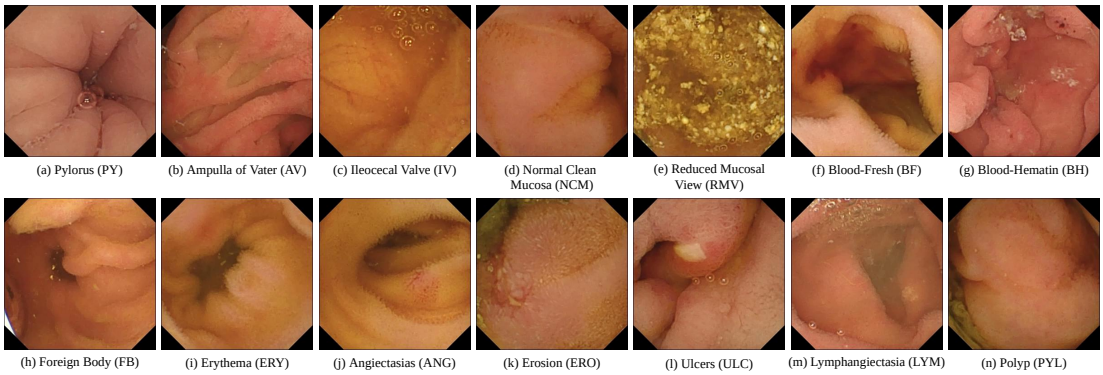


Figure 3.1: *The frames extracted from reference videos in the QVCED dataset represent a diverse range of findings.*

following subsection provides a comprehensive description and the selection process of the reference videos.

2.1 Reference Videos Selection

The selection of the reference videos aimed at optimizing a wide range of pathological scenarios and maximizing continuous temporal information which enables a thorough evaluation and analysis of quality enhancement algorithms, facilitating advancements in the WCE domain.

To ensure scene content diversity, the QVCED dataset includes fourteen distinct categories. These categories encompass various WCE findings such as Pylorus (PY), Ampulla of Vater (AV), Ileocecal Valve (IV), Normal Clean Mucosa (NCM), Reduced Mucosal View (RMV), Blood-Fresh (BF), Blood-Hematin (BH), Foreign Body (FB), Erythema (ERY), Angiectasias (ANG), Erosion (ERO), Ulcers (ULC), Lymphangiectasia (LYM), and Polyp (PYL). This diverse selection allows for a comprehensive evaluation of algorithms and techniques in the field of WCE, covering a broad spectrum of medical scenarios commonly encountered in clinical practice.

Some images from reference videos are shown in Fig. 3.1. In the following subsection, the IQA metrics which are used to evaluate the quality of each video are presented. These metrics help identify the highest-quality videos for each finding, ensuring that the chosen reference videos meet the required quality standards which enhances the dataset's reliability and usefulness for various research purposes.

2.1.1 Noise Assessment

To estimate the noise level, we employ the fast noise variance estimator proposed by Immerkaer [239]. The process begins by converting the input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ into grayscale, resulting in the grayscale image $\mathbf{I}_{gray} \in \mathbb{R}^{H \times W}$ where H, W are the height and width of the images. A noise estimation mask, denoted as $\mathbf{M} \in \mathbb{R}^{3 \times 3}$, is then used to estimate the standard deviation of additive white Gaussian noise in the image. This mask is derived from two elements approximating the Laplacian of the image. The estimated standard deviation of the noise $\hat{\sigma}_n$ is computed as follows:

$$\hat{\sigma}_n = \sqrt{\frac{\pi}{2}} \frac{\Sigma_{x,y} |\mathbf{I}_{gray}(x,y) * \mathbf{M}|}{6(W-2)(H-2)}, \tag{3.1}$$

where $\mathbf{M} = 2(\mathbf{L}_2 - \mathbf{L}_1)$ with the given $\mathbf{L}_1, \mathbf{L}_2$:

$$\mathbf{L}_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{3.2}$$

$$\mathbf{L}_2 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{bmatrix} \tag{3.3}$$

The noise variance σ_n^2 could be estimated, as in [240], using the robust median estimator of the highest sub-band of a Daubechies Two wavelet transform:

$$\sigma_n^2 = \frac{Median(HH)}{0.6745}, \tag{3.4}$$

where HH are the high-high band wavelet coefficients. An alternative approach is to let $\sigma_n^2 = Variance(HH)$.

2.1.2 Blur Assessment

To measure the level of blur in an image, the Perceptual Blur Index (PBI) [241] was used as a thresholding metric. The PBI metric takes into account the perceptual differences in how the Human Visual System (HVS) perceives the addition of blur to an already blurred image compared to a sharp one. Mathematically, the PBI is defined as the difference between the total radial energy of the input image, denoted as $ER(w)$, and the total radial energy of its binomial filtered version, denoted as $ER_f(w)$. The formula for calculating the PBI is as:

$$PBI = log\left(\frac{1}{w_{max}} \Sigma_w |ER(w) - ER_f(w)|\right) \tag{3.5}$$

$$ER(w) = \frac{1}{K} \Sigma_K |F(w, \theta_k)|, \theta_k = \frac{k\pi}{K}, \tag{3.6}$$

$$ER_f(w) = \frac{1}{K}\Sigma_K |F_f(w,\theta_k)|, \theta_k = \frac{k\pi}{K}, \tag{3.7}$$

where $F(w,\theta_k)$ and $F_f(w,\theta_k)$ represent the centered Fourier coefficients of the input images and its binomial filtered version, respectively, in the polar coordinates.

2.1.3 Uneven Illumination Assessment

To assess the presence of uneven illumination, the Illumination Histogram Equalization Difference (IHED) [242] is employed. IHED measures the impact of histogram equalization (HE) on the spatial distribution of background illuminance (BI). The evaluation process involves converting the image into the HSV color space to extract the brightness channel $\mathbf{V} \in \mathbb{R}^{H\times W}$. Subsequently, the background illuminance $\mathbf{BI}(x,y) \in \mathbb{R}^{H\times W}$ is extracted through the application of a low-pass filtering method of size $h = \frac{H}{4}$. Finally, IHED is calculated using the following formula:

$$IHED = \frac{\sigma_D}{\frac{1}{H\times W}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\mathbf{BI}(x,y)}, \tag{3.8}$$

where σ_D is the standard deviation of the difference signal $\mathbf{D}$, which is computed as:

$$\mathbf{D}(x,y) = \mid \mathbf{BI}(x,y) - \mathcal{T}(\mathbf{BI}(x,y)) \mid, \tag{3.9}$$

where $\mathcal{T}$ denotes the histogram equalization transformation.

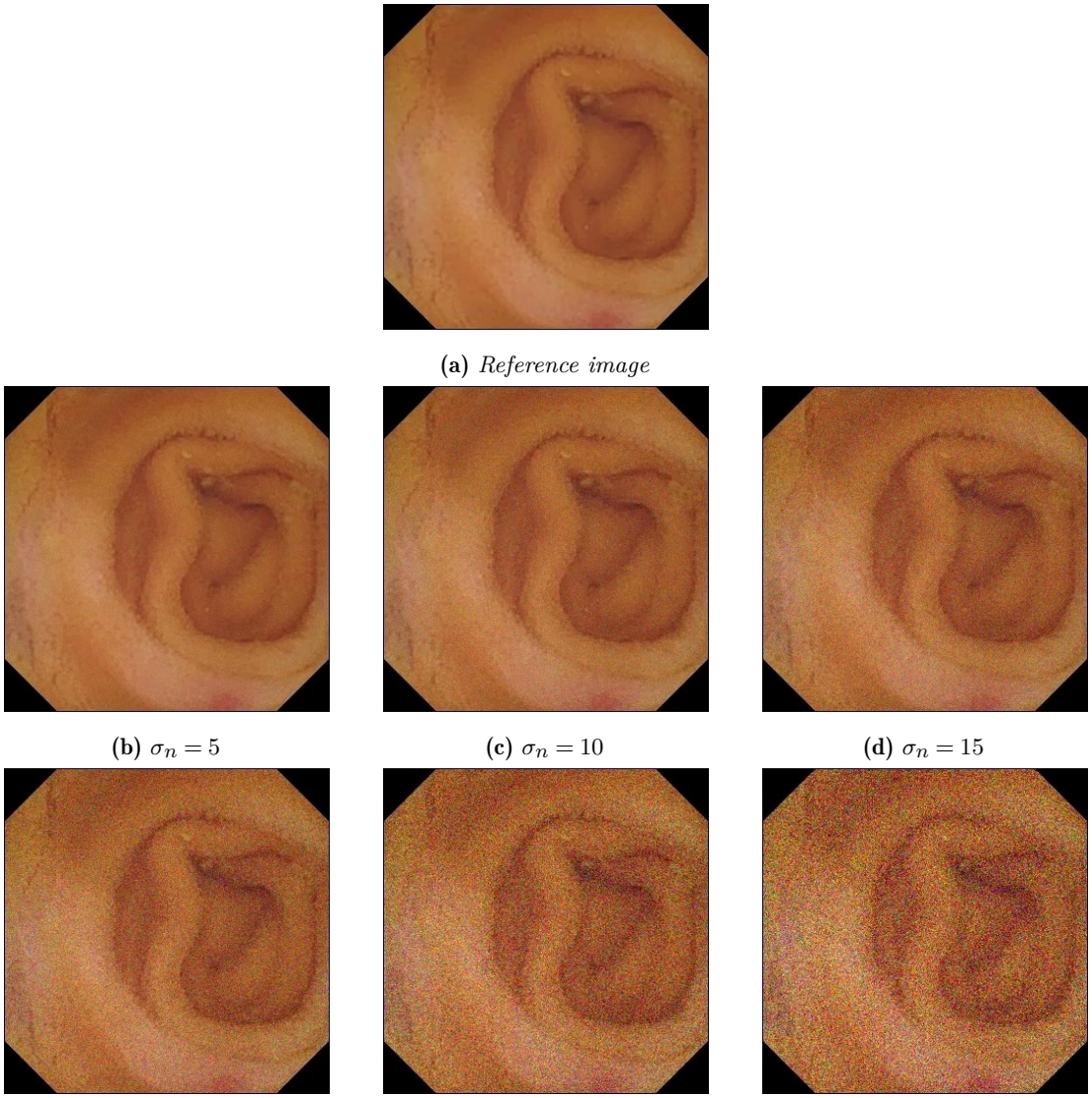
A video is considered acceptable for use as a reference only if the levels of all three distortions (i.e., noise, blur, and uneven illumination) are below a predetermined threshold. Once the reference video is chosen, the next subsection will outline how we simulated distortions on these selected reference videos.

2.2 Distortion Generation

We have integrated four prevalent degradations consisting of noise, uneven illumination, defocus, and motion blur into our extensive dataset. To ensure a faithful reproduction of each distortion, we have applied suitable mathematical models to every individual frame of the reference video. In the current stage of our research, we only added one type of distortion to each video, with the same severity throughout its entirety.

2.2.1 Noise Model

Noise is a common distortion in video systems, particularly in WCE. It is caused by narrow apertures, small sensors, and limited dynamic range [19] and negatively impacts the effectiveness of the endoscopic examination process. In our study, we included the



(a) *Reference image*

(b) $\sigma_n = 5$ **(c)** $\sigma_n = 10$ **(d)** $\sigma_n = 15$

(e) $\sigma_n = 20$ **(f)** $\sigma_n = 30$ **(g)** $\sigma_n = 40$

Figure 3.2: *The results of adding AWGN into the reference image with 6 different levels.*

Additive White Gaussian Noise (AWGN) model in our dataset which assumes that the video noise follows a Gaussian distribution. Mathematically, the distorted image can be represented as:

$$\mathbf{I}_{noisy} = \mathbf{I} + \mathbf{N}, \tag{3.10}$$

where $\mathbf{I}$ represents the original image, and $\mathbf{N} \sim \mathcal{N}(0, \sigma_n^2)$ represents the random noise value following a Gaussian distribution with standard deviation σ_n. To control the severity, AWGN level is configured with $\sigma_n \in \{5, 10, 20, 30\}$. These various noise levels are visually demonstrated in Fig. 3.2.

2.2.2 Defocus Blur Model

In WCE, the wireless capsule is equipped with a fixed-focus lens endoscope [243]. This design introduces defocus blur when objects in the scene are not precisely at the camera's focal distance. To simulate defocus blur, a low-pass filtering of the input image using an isotropic Gaussian impulse response as shown in Fig. 3.4a is commonly used. The isotropic Gaussian kernel is used to simulate the rotational symmetry around the optical axis of the blurring effect. The impulse response associated with this blur, denoted as $h_{db}(x,y)$, is defined as:

$$h_{db}(x,y) = \frac{1}{2\pi\sigma_{db}^2} \exp \frac{-(x^2+y^2)}{2\sigma_{db}^2}, \ \sigma_{db} \in \{1,2,3,5\} \tag{3.11}$$

The blurring extent is determined by the standard deviation parameter σ_{db}. Increasing σ_{db} leads to stronger smoothing and more noticeable blurring effects. The size of the convolution mask, denoted as W_{db}, is chosen to preserve the energy of the filtered image signal. To preserve 99% of the total energy of the Gaussian, a size of $6\sigma_{db}$ at least is required. The filter size W_{db} should be an odd number as:

$$W_{db} = 2 \times \lceil 3\sigma_{db} \rceil + 1 \tag{3.12}$$

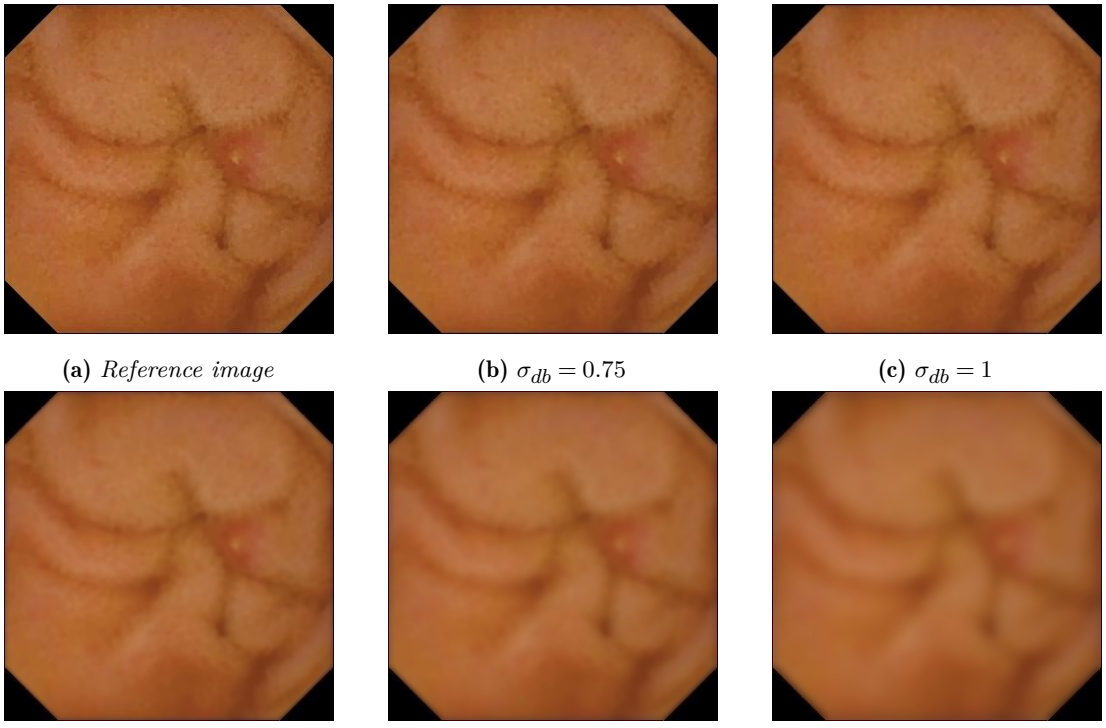
Fig. 3.4a illustrates an example of a defocus blur kernel of standard deviation $\sigma_{db} = 1$. Fig. 3.3 shows us the results of adding defocus blur.

2.2.3 Motion Blur Model

The rapid and sudden movements of the capsule endoscope can cause blurring, influenced by factors such as fast camera motions at low frame rates, the inability to adjust lens focus, camera mechanism instability, and sensor sensitivity to light variations [243]. When the capsule endoscope moves in a straight line, it results in linear motion blur. The blur kernel, denoted as h_{mb}, can be formulated using two known parameters: the direction of motion blur θ_{mb} and the length of motion blur $L_{mb} \in \{5,10,15,25\}$. The formulation is as follows:

$$h_{mb}(x,y) = \begin{cases} \frac{1}{L_{mb}}, & \text{if } \sqrt{x^2+y^2} \leq \frac{L_{mb}}{2}, -\tan\theta_{mb} = \frac{x}{y}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.13}$$

We evaluate the motion in each frame in order to add a motion blur corresponding to the actual motion of the video. In the initial stages, the Lucas-Kanade method [244] is commonly used for estimating the movement direction of a capsule endoscope through optical flow estimation. Two consecutive frames captured by the capsule endoscope are subtracted and the Otsu thresholding technique [245] is applied to generate a foreground

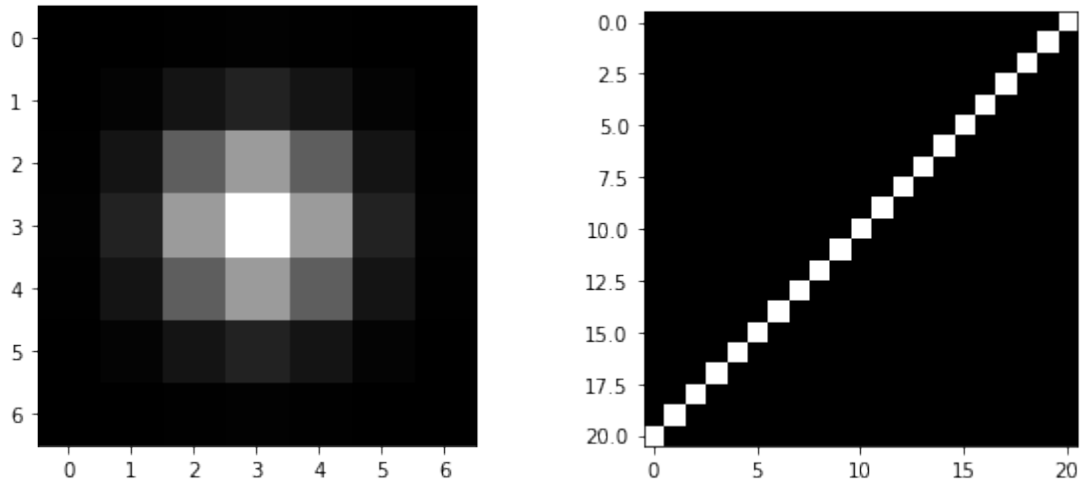


(a) *Reference image* **(b)** $\sigma_{db} = 0.75$ **(c)** $\sigma_{db} = 1$

(d) $\sigma_{db} = 2$ **(e)** $\sigma_{db} = 3$ **(f)** $\sigma_{db} = 5$

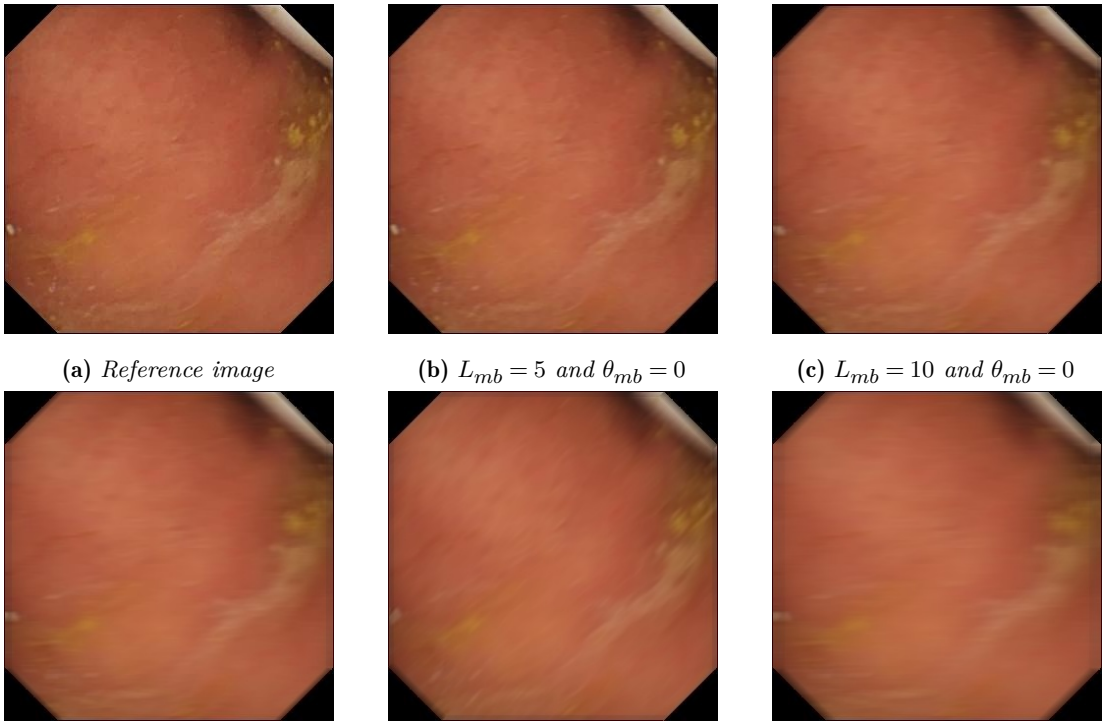
Figure 3.3: *The results of adding defocus blur into the reference image with 5 different levels.*

binary map. The optical flow is then estimated using the Lucas-Kanade method on the center of gravity of the foreground. Fig. 3.4b illustrates an example of a motion blur kernel, which is configured by two parameters: the direction θ_{mb} and the length L_{mb}. Fig. 3.5 shows us the results of adding motion blur.



(a) *Defocus blur kernel* **(b)** *Motion blur kernel*

Figure 3.4: *Visual representation of the blurring kernel, characterized by a defocus blur standard deviation of* $\sigma_{db} = 1$ *and motion blur with* $L_{mb} = 20, \theta_{mb} = \frac{\pi}{4}$*, respectively.*



(a) *Reference image* **(b)** $L_{mb} = 5$ *and* $\theta_{mb} = 0$ **(c)** $L_{mb} = 10$ *and* $\theta_{mb} = 0$

(d) $L_{mb} = 15$ *and* $\theta_{mb} = 0$ **(e)** $L_{mb} = 15$ *and* $\theta_{mb} = 45$ **(f)** $L_{mb} = 25$ *and* $\theta_{mb} = 0$

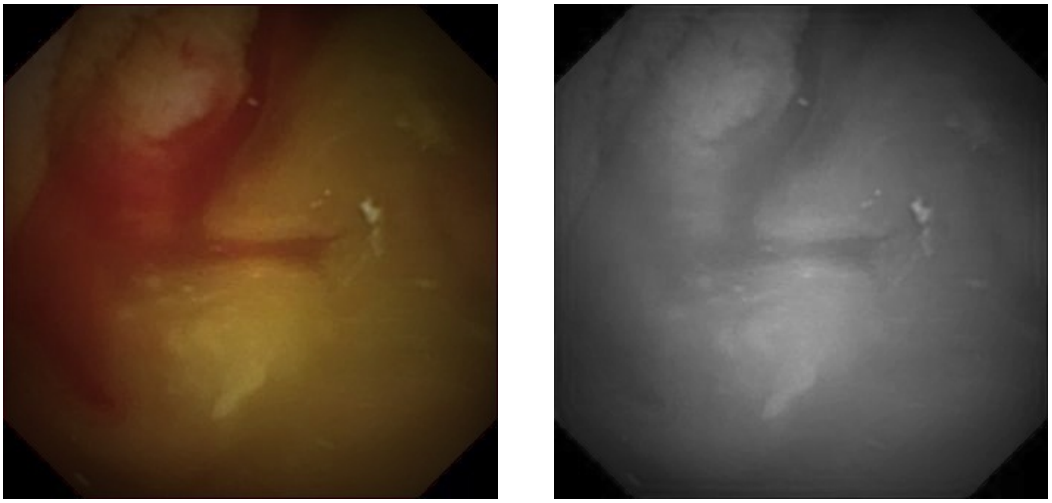
Figure 3.5: *The results of adding motion blur into the reference image with 4 different levels. Fig. (d)-(e) shows us an example of changing the direction from 0 degrees to 45 degrees.*

2.2.4 Uneven Illumination Model

The motion of the capsule endoscope, caused by the gastrointestinal tract's peristaltic activity and limited capsule light, can introduce uneven illumination. To simulate this effect, the reference image is first converted from the RGB color space to the HSV color spaces as shown in Fig. 3.6. Subsequently, we perform a pointwise multiplication of the reference image with a mask. The coefficients of this mask, which are determined in the spatial plane (Fig. 3.7a), can be represented mathematically as a hybrid distribution that integrates two-dimensional distributions from one normal variable and one log-normal variation[246] as:

$$p(\hat{\mathbf{x}} \mid \mu, \Sigma) = \left(\prod_{i=p+1}^{N} \frac{1}{x_i} \right) \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \times \exp\left(-\frac{1}{2} (\hat{\mathbf{x}} - \mu)^T \Sigma^{-1} (\hat{\mathbf{x}} - \mu) \right), \tag{3.14}$$

where p, q are the normal and log-normal variants, respectively. μ is the mean vector, Σ is the (symmetric, positive definite) covariance matrix (of size $N \times N$), and $|\Sigma|$ is its determinant. N is the total dimensions of two variants, $\hat{\mathbf{x}}^T = (x_p \; ln(x_q))$, μ is the



(a) *Original image* **(b)** *V-channel of image*

Figure 3.6: *The results of converting the image from RGB to HSV color space.*

vector of means.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix} \tag{3.15}$$

where $\rho_{i,j}$ are the correlations between the variables, σ_j is the associated standard deviation for $\hat{x}_j$. Fig. 3.7b,c show the generated masks of the hybrid distribution in two different angles θ, respectively. However, in this preliminary work, only a simulated circular-gradient mask was taken into account. As depicted in Fig. 3.8, the mask $\mathbf{M}(x,y) \in \mathbb{R}^{H\times W}$ is defined to conform to the dimensions of the original image.

$$\mathbf{M}(x,y) = 255 - \left[\frac{2\Delta_I}{W}\sqrt{(x-x_c)^2+(y-y_c)^2}\right], \tag{3.16}$$

where $\mathbf{M}(x_c,y_c)$ is the circle center at coordinates $(x_c,y_c) \in \{(112,224),(168,168),(224,224)\}$. To achieve varying levels of illumination, the difference in intensity between the brightest pixel of the image and the darkest pixel is set as $\Delta_I \in \{80,100,135,170\}$. In the future, we plan on using the previously defined hybrid distribution as a mask.

With uneven illumination, 4 different levels are also considered. Besides, 3 positions (center, top-right, and bottom-right) are used to adjust the center of the circle gradient. The level of the gradient is controlled by adjusting the Δ_I (Eqs. (3.16)), where $W = 336$ is the size of images in the Kvasir-Capsule dataset. The following Fig. 3.9 shows us the results.

In summary, the dataset creation process involved the simulation of four distinct types of distortion, each type having four different severity levels. This resulted in a total of 320 degraded videos in the QVCED dataset. A comprehensive overview of the

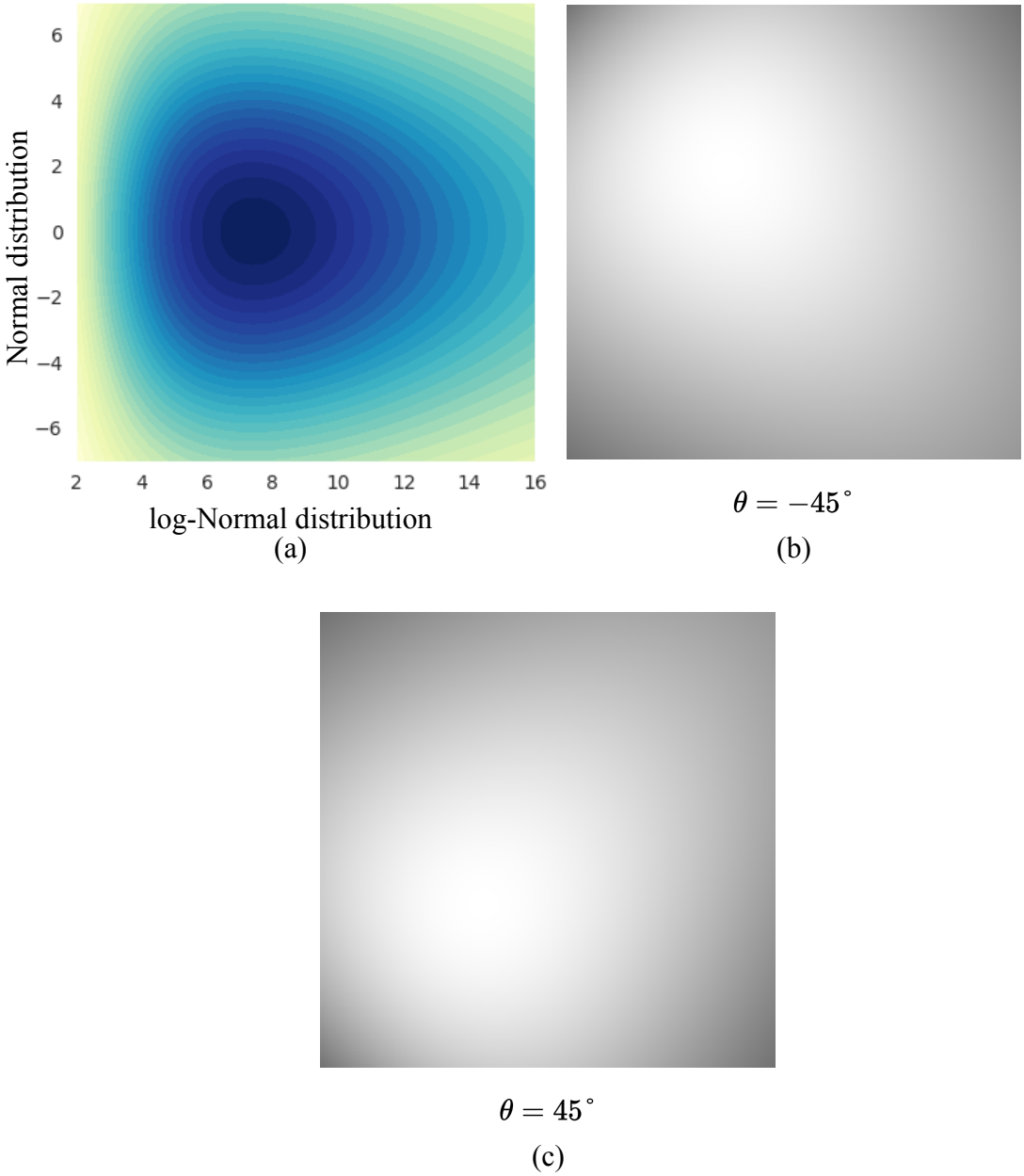
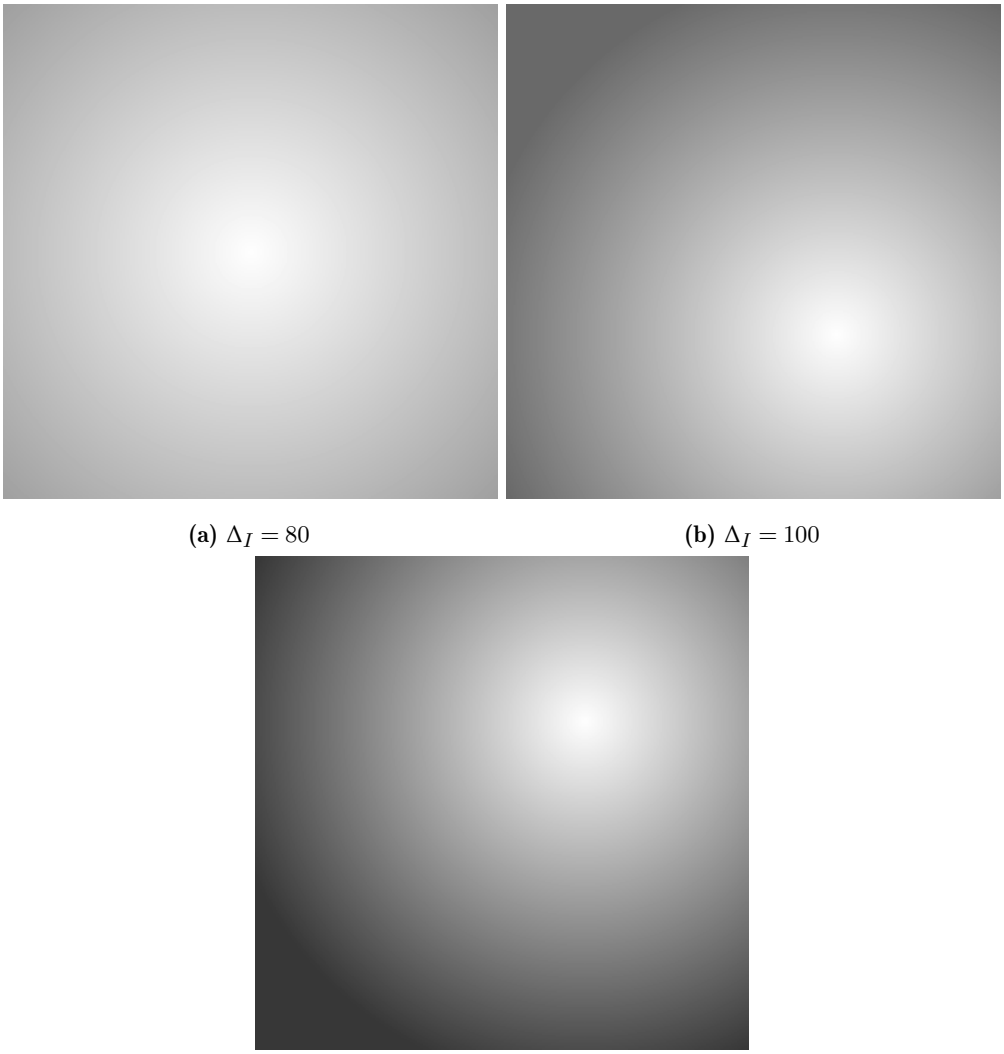


Figure 3.7: *Gradient masks to simulate the Uneven Illumination which is represented mathematically as a hybrid distribution.*

dataset, including specific details such as the types of distortions and severity levels, can be found in Table 3.1.

In the following section, some extensive experiments are conducted to analyze the diversity and practical applicability of the proposed dataset.



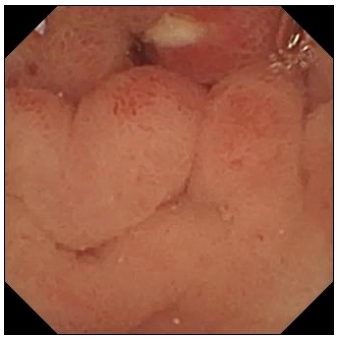
(a) $\Delta_I = 80$ **(b)** $\Delta_I = 100$

(c) $\Delta_I = 135$

Figure 3.8: *Some examples of the artificial circular-gradient Uneven Illumination mask.*

3 Dataset Analysis

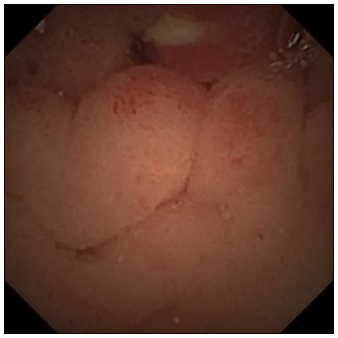
To evaluate the proposed dataset, some experimental studies were conducted. First, a subjective test (Section 3.1) was performed to assess and validate the dataset's quality and its alignment with human perception. This test evaluates how well the dataset is perceived by human observers and ensures its overall quality. In a second round of experiments, the statistical features of the dataset were analyzed to verify its content diversity (Section 3.2).



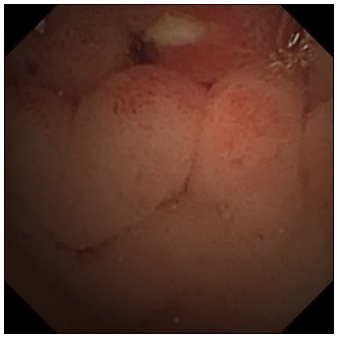
(a) *Reference image*

(b) $\Delta_I = 80$ *and the position is center*

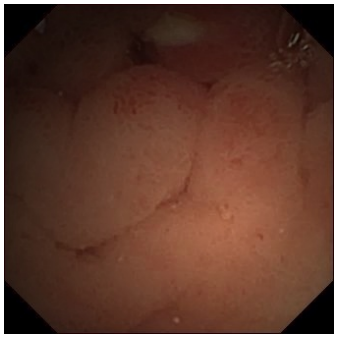
(c) $\Delta_I = 100$ *and the position is center*



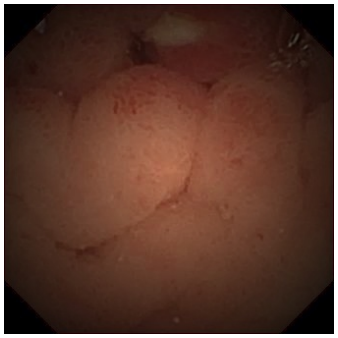
(d) $\Delta_I = 135$ *and the position is center*



(e) $\Delta_I = 135$ *and the position is top-right*



(f) $\Delta_I = 135$ *and the position is bottom-right*



(g) $\Delta_I = 170$ *and the position is center*

Figure 3.9: *The results of adding uneven illumination into the reference image with 4 different levels and 3 possible positions.*

Table 3.1: *Summary of the proposed wireless capsule endoscopy video quality assessment dataset.*

Number of Reference Videos	20	**Number of Distorted Videos**	320
Number of Findings	14	**Pathologies**	6
Resolution of Videos	336×336	**Frame Rate**	30
Duration	10s	**Video Type**	.mp4
Number of Distortions	4	**Level of Distortion**	4
Distortion Types	Noise, Defocus Blur, Motion Blur, Uneven Illumination		

3.1 Subjective Test

Prior to the main WCE subjective test, observers underwent the Ishihara 38 plates CVD verification[247] to detect any red-green color deficiencies. Participants with an accuracy above 70% were selected to complete the WCE subjective test, ensuring normal color perception for accurate evaluation.

To conduct the WCE subjective testing process efficiently, a pair comparison protocol based on the ITU-T standard[248] was implemented, following the described testing environment.

3.1.1 Testing Environment

In the WCE quality assessment subjective test, observers were presented with randomized pairs of distorted videos and corresponding reference videos. Randomization was implemented to eliminate presentation order bias. For each video pair, observers were asked to provide an opinion score indicating the perceived severity of distortion. The implemented four-point scale corresponds to four distortion severity levels including: (1) Hardly Visible, (2) Just Noticeable, (3) Annoying, and (4) Very Annoying. The obtained opinion scores allowed us to assess the subjective quality of the distorted videos compared to the corresponding reference videos. The Mean Opinion Score (MOS) for a video is the average score given by observers for that video.

An online platform (shown in Fig. 3.10) was developed and designed to facilitate the conduct of subjective tests. The platform underwent thorough optimization to

ensure usability, including aspects such as background color, button size, and position. These optimizations aimed to enhance the testing experience and make it convenient for observers to effectively complete the task following the ITU-T standard. The source code of the platform is available at: https://github.com/tansyab1/WCETest.

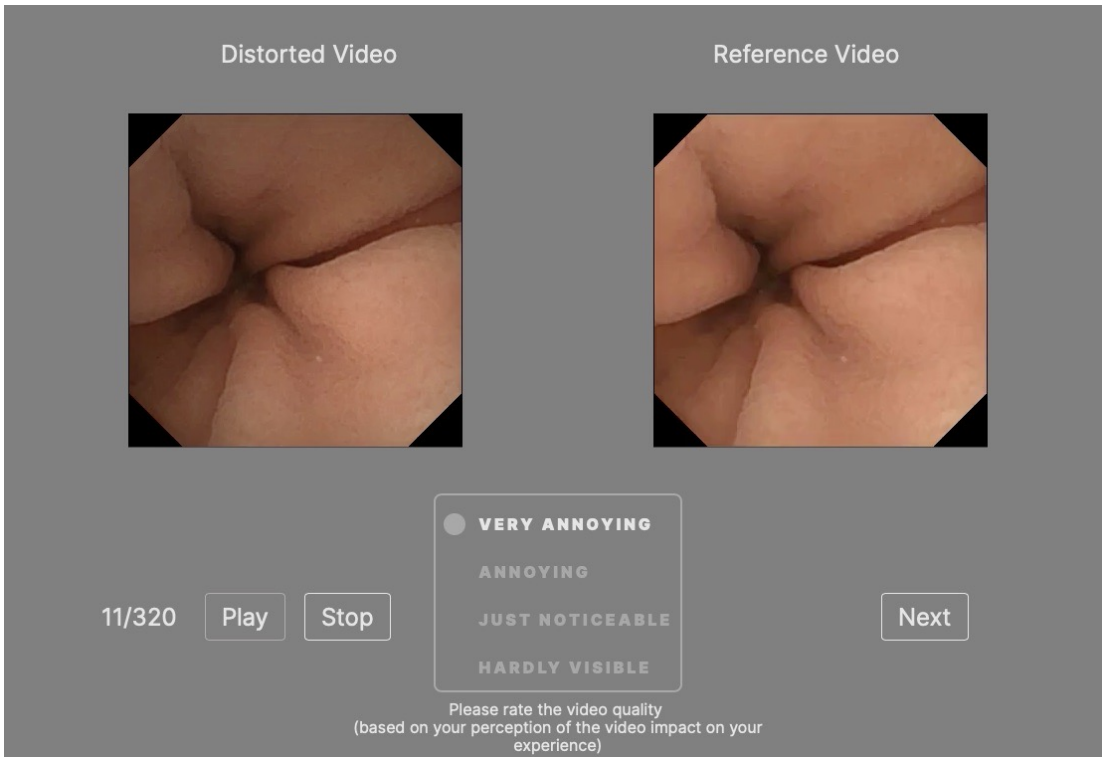
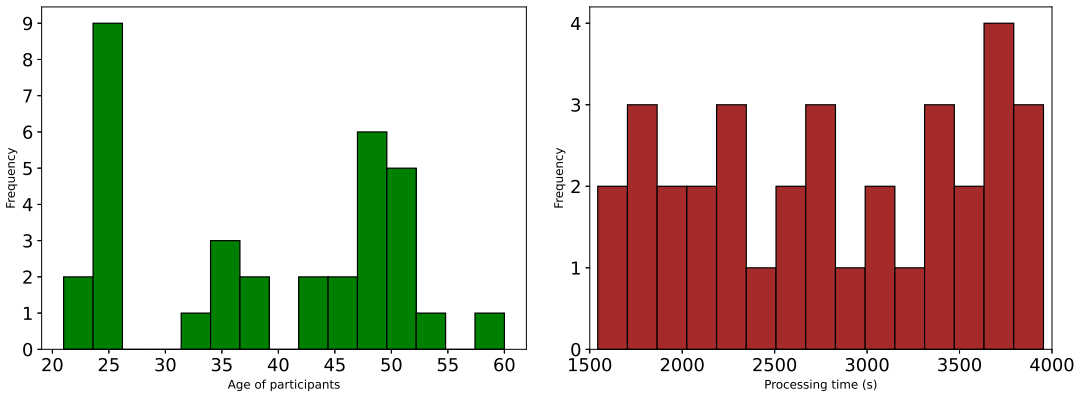


Figure 3.10: *WCE subjective test window.*

A total of 34 individuals, comprising 12 experts and 22 non-experts, with diverse age groups and backgrounds, took part in subjective the experiments. Fig. 3.11 displays the distribution of participants' age and the duration of their involvement in the subjective test. Participants across various age groups were included in the subjective test, as shown in Fig. 3.11a. This diverse age distribution enhances the reliability of the test outcomes by avoiding biases toward any specific age category. Moreover, it is clearly noticeable from Fig. 3.11b that each participant dedicated a minimum of approximately 5 seconds to evaluate each video, demonstrating their focused attention and commitment to efficient testing. This statistic affirms the credibility and applicability of the test's outcome.

3.1.2 Video Quality Score

As mentioned earlier, the evaluation includes four levels of distortion. Level 1 represents the minimal degradation of distortion, while level 4 represents the most severe condition,



(a) *Age* **(b)** *Processing time (s)*

Figure 3.11: *Age and the processing time distributions of observers participated in the subjective experiment.*

where a higher value indicates a lower-quality perception for the video observer. Fig. 3.12

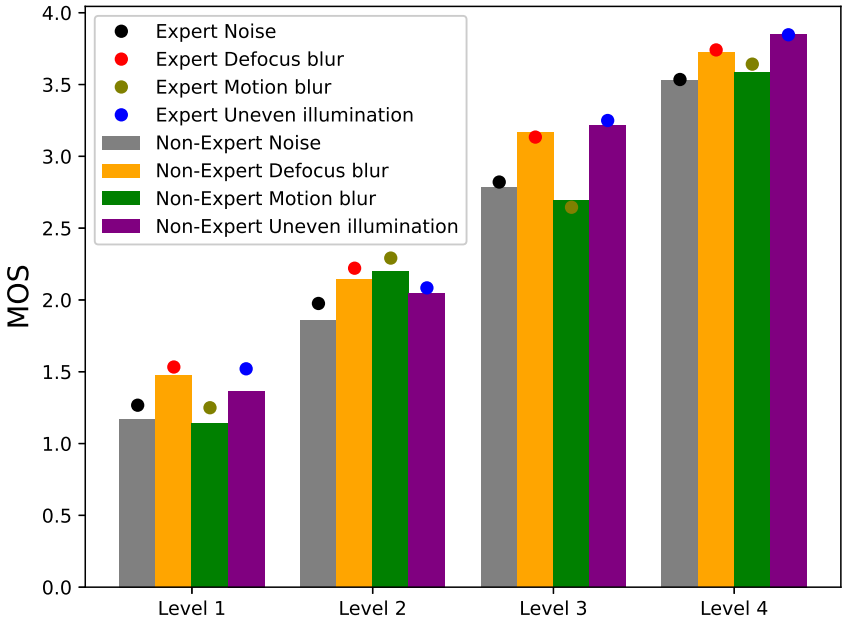


Figure 3.12: *Comparison of the subjective score regarding experts and non-experts.*

compares the mean scores of experts and non-experts for the proposed dataset. The presented data illustrates that experts and non-experts hold a significant correlation between the obtained scores. However, the experts exhibit a heightened level of attention toward specific tasks, which enhances their sensitivity to even the slightest deviations. Therefore, the dissimilarity is more conspicuous when examining videos exhibiting low levels of distortion.

3.2 Diversity Data Analysis

To analyze the dataset's diversity and broad applicability, experiments were conducted to verify significant variations in video content. These experiments provide valuable insights and benefits for various image-processing tasks. A diverse dataset serves as a valuable resource for benchmarking, validation, and training, facilitating significant advancements in image processing.

To evaluate the content diversity of the datasets, we used deep features of dimension 4096 extracted from a pre-trained VGG-16[249] on ImageNet[250]. By employing t-SNE (t-distributed Stochastic Neighbor Embedding) [251], we projected these high-dimensional features onto a 2D subspace. The resulting visualization, shown in Fig. 3.13, succinctly represents the content diversity across the datasets. The broad spectrum displayed in the visualization illustrates the extensive range and variety of visual content present in the dataset. In addition to the primary analysis, a further examination of

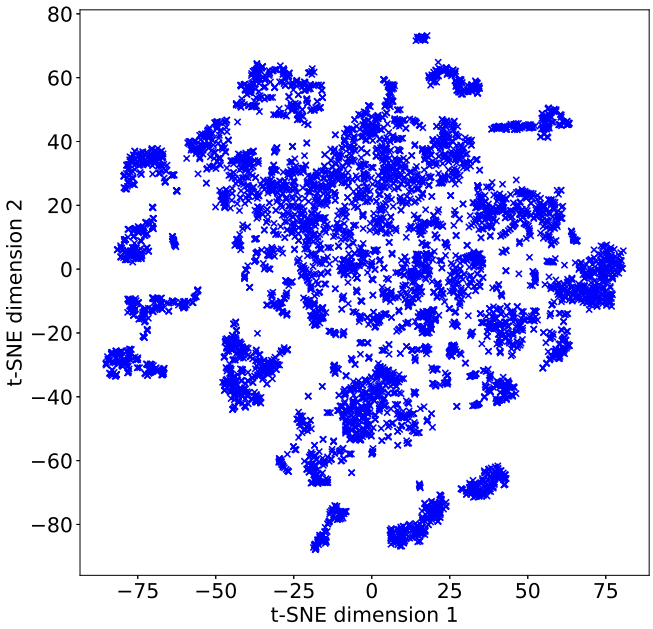
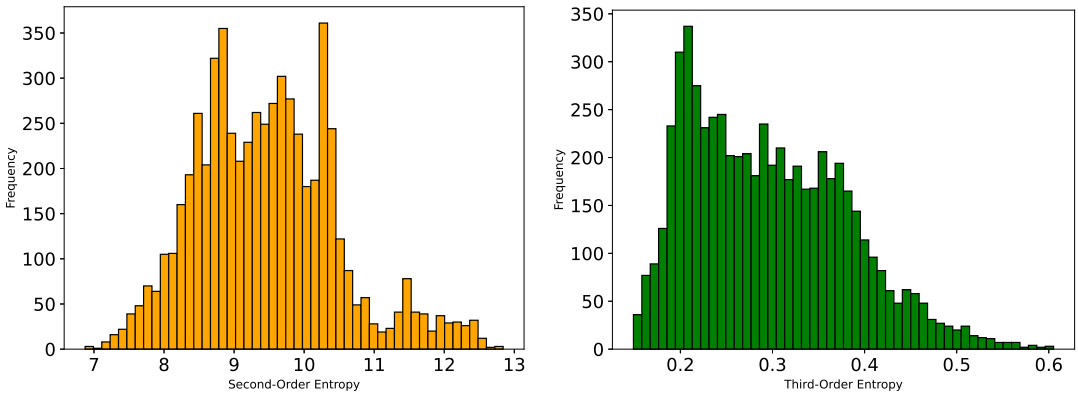


Figure 3.13: *t-SNE visualization of the embedded feature generated from 20 reference videos by VGG-16 pre-trained network.*

the dataset is conducted to analyze the distribution of image entropies, considering both spatial and temporal information. Second-order entropy analysis provides insights into the spatial features within the dataset, where higher entropy indicates a greater content diversity of images. Furthermore, the incorporation of third-order entropy analysis takes into account inter-frame information. By considering the relationships and dependencies between consecutive frames, the third-order entropy provides a deeper

understanding of the temporal dynamics and variations within the dataset. This analysis offers a comprehensive perspective on the dataset's complexity and richness. In this work, calculations were performed on a dataset of 6000 images from 20 reference videos. Fig. 3.14 illustrates the broad histogram of the entropies, indicating a wide range of visual features and affirming the practical applicability, diversity, and usefulness of the QVCED dataset.



(a) *Second-Order entropy* **(b)** *Third-Order entropy*

Figure 3.14: *Distribution of Variations and Entropy among 20 reference videos.*

4 Conclusion

In this chapter, we have introduced a preliminary quality assessment dataset specifically designed for wireless capsule endoscopy. QVCED comprises four distinct distortion types, with each type further subdivided into four levels, resulting in a total of sixteen variations. This dataset serves as a quality assessment resource specifically targeting the WCE domain. Especially, it addresses the previously neglected data challenge and offers valuable insights for evaluating and analyzing the effectiveness of image and video processing algorithms in this particular field. The dataset's strength lies in the extensive diversity of its visual content, enabling researchers to tackle demanding real-world contexts. In this work, the addition of synthetic distortion to a given frame may not have a noticeable impact if the frame is already affected by authentic distortion. In the future, we could remove any existing distortion before applying a synthetic one. To accurately identify and address distortion in WCE images, it is imperative to have access to a comprehensive dataset that includes information about both the type and level of distortion present. In the upcoming chapter, we will present our proposed Image Quality Assessment (IQA) metric, designed specifically to evaluate the severity of distortions, with a particular focus on addressing issues related to uneven illumination.

CHAPTER 4

The proposed IQA metric: *A No-Reference Measure for Uneven Illumination Assessment on Laparoscopy and WCE images*

Abstract

A frequent degradation in video-guided surgery especially laparoscopic and in WCE is uneven illumination, due in large part to physical limitations of the sensors and uncontrolled lighting conditions in the internal structure of the digestive tract and particularly at the level of the intestines. Surgical as well as postoperative task accuracy can be seriously affected by the perceptual quality of the acquired images or videos. In this respect, a No-Reference Image Quality Assessment (NR-IQA) metric dedicated to uneven illumination is proposed in this work. The key idea is to analyze the effect of contrast enhancement on the spatial distribution of the luminance component of the signal. The results obtained through extensive experiments, performed on dedicated databases of different medical imaging modalities including laparoscopic images and WCE images, have shown that the proposed metric significantly improves the state-of-the-art NR-IQA metrics when applied for uneven illumination assessment.[1]

[1]T. -S. Nguyen, J. Chaussard, M. Luong, H. Zaag and A. Beghdadi, *"A No-Reference Measure for Uneven Illumination Assessment on Laparoscopic Images,"* 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 4103-4107, doi: 10.1109/ICIP46576.2022.9897302.

Chapter content

1 Introduction 76

2 Proposed Method 82

2.1 Color Space Conversion 82

2.2 Background Illuminance Estimation 83

2.3 Uneven Illumination Assessment 84

3 Performance Evaluation on laparoscopic images 85

3.1 Dataset 86

3.2 Correlation with Human Quality Judgment 86

4 Performance Evaluation on WCE images 87

4.1 Correlation with Human Quality Judgment 87

5 Conclusion 89

1 Introduction

In the previous chapter, we mentioned about the image quality and its pivotal role in influencing the performance of classification tasks. Indeed, the acquisition condition of video signals and in particular the problem of non-uniform lighting, specular reflection and noise due to physical limitations of sensors generate a visual discomfort for not only the surgeon, which negatively affects the performance of surgery (e.g., laparoscopy) and diagnosis (e.g., WCE). This consequently increases the risk of misdiagnosis and uncontrolled surgical procedures with fatal and irreversible consequences [252].

More particularly, the acquired videos often suffer from uneven illumination, e.g., poor illumination not only in the peripheral areas of the light spot but also in the weakly illuminated regions due to the relief and texture of the gastrointestinal tissues [253]. Uneven illumination (UI) manifests itself as a non-uniform spatial distribution of light, making certain regions appear more illuminated than others, regardless of the intensity of the incident light beam. This may affect the visibility and discrimination of the subtle features of vital roles [254]. Moreover, the performance of some subsequent video processing and analysis tasks, e.g. segmentation or 3D reconstruction may be negatively affected by this photometric distortion [255], [256]. It is then important to control this distortion by measuring it accurately at the acquisition level in order to design an effective image quality enhancement.

Therefore, an image quality assessment (IQA) which allows to objectively quantify the levels of the illumination non-uniformity is essential for the performance of any video-guided surgery or diagnosis system whose accuracy and reliability highly depend on the quality of the acquired images.

While subjective assessment of image quality is a very reliable way to evaluate image quality which can be used for benchmarking objective quality metrics, it is complex, time-consuming and generally requires a controlled environment that should meet ITU standard requirements [25]. Furthermore, subjective evaluation cannot be introduced in real applications that require real-time responses and decisions. This has led to the development of objective image quality metrics. Objective IQA methods can be subdivided into three categories: Full-Reference IQA (FR-IQA) [26], Reduced-Reference IQA (RR-IQA)[27] and No-Reference or Blind IQA (NR-IQA)[28]–[30]. In practice, pristine reference images are not always available [73]. Therefore, searching for NR-IQA measures is the most practical solution to meet real needs in many real-world applications. NR-IQA refers to the automatic prediction of the quality of a distorted image using its inherent features or/and prior knowledge about the distortion through

some more or less simple image formation models.

It is worth noticing that there are also some interesting NR-IQA approaches that do not take into account the distortion-specific features and are therefore able to predict the image quality in the case of multiple distortions [28]–[30]. In particular, Mittal et al. [28] proposed the Natural Image Quality Evaluator (NIQE) metric based on a natural scene statistic model.

It should be noted that several metrics have been proposed to estimate the objective image quality affected by different types of distortions such as additive noise, blurring or compression artifacts [58], [91]. However, only a few NR metrics were proposed for uneven illumination assessment and more particularly in the context of medical imaging [92], [93]. An interesting NR-IQA metric, called Average Gradient of the Illumination Component (AGIC), developed for Dermoscopy Images has been introduced in [92]. It is based on the average gradient of the illumination component estimated by using a Weber Law-based gradient measure. Firstly, the image is divided into $M \times N$ small rectangle patches. For patch (i,j), the gradient is defined as the max difference between patch (i,j) with its neighbor patches.

$$g(i,j) = \max|h(i,j) - h_k(i,j)|, k = 1,2,\ldots 8 \tag{4.1}$$

where $i \in \{1,2,\ldots M\}$, and $j \in \{1,2,\ldots N\}$,$h(i,j)$ and $h_k(i,j)$ represent the average gray value of patch (i,j) and its 8-connected neighbors respectively. There is a fact that a whispered voice in a quiet room is easily caught, but shouting in a noisy condition may not be noticeable [93].Weber's law is approximately true for the perception of light intensity, that is, the perceived light intensity change is relative to the background light intensity. Taking Weber's law into consideration, the new gradient is defined as:

$$g_w(i,j) = \frac{\max|h(i,j) - h_k(i,j)|}{h(i,j)} \tag{4.2}$$

Then the average gradient AGIC is defined as:

$$AGIC = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} g_w(i,j) \tag{4.3}$$

A higher AGIC value is indicative of more pronounced uneven illumination within an image. However, it's important to note that this metric evaluates the image primarily based on the average gradient level of the illumination background (IB). Consequently, in cases where the primary area of the image is dark and there is minimal contrast or change between the dark region and the main portion of the image, this criterion can yield inaccurate results.

To illustrate this point, the subsequent Fig. 4.1 presents the outcomes of testing AGIC on a selection of videos characterized by varying degrees of uneven illumination. For a more in-depth analysis, let's consider that as the difference Δ_v between the highest

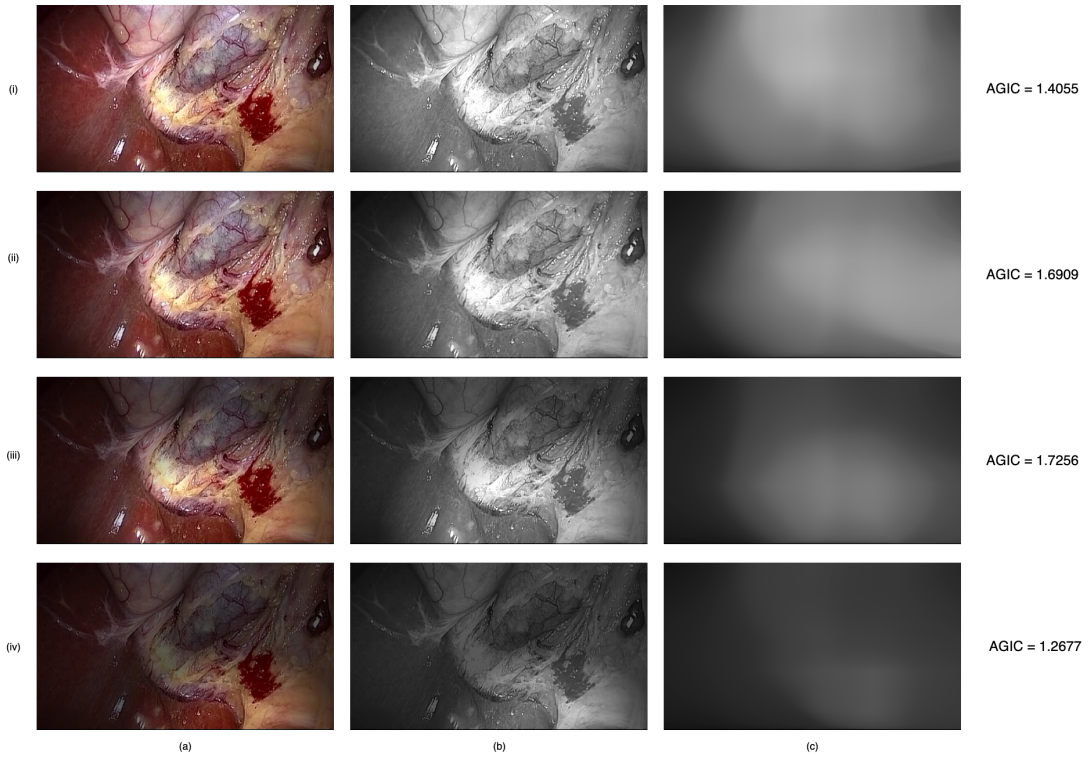


Figure 4.1: *The AGIC result on the image distorted by 4 levels of uneven illumination. The experiment was operated on LVQ dataset [257]. The column (a) is the original image I. Column (b) is the extracted V-channel V from the original image I. The column (c) is the IB of V which is extracted using LPF ($X = LPF(V)$). The fourth column is the corresponding AGIC value of each IB.*

illuminance v_{center} and the lowest illuminance v_{border} grows larger, the gradient value between the bright region and the dark region will also increase. Consequently, this escalation in gradient values signifies a corresponding increase in the degree of uneven illumination. As depicted in Fig. 4.2, it becomes evident that the AGIC value in the last scenario deviates significantly from what one might expect. This behavior mirrors the observations made in real images from the LVQ dataset [257]. The findings derived from these synthesized images lead us to the conclusion that AGIC may misinterpret and erroneously categorize images with good overall quality and low-light backgrounds. This misclassification occurs because AGIC relies on the assessment of the image's average gradient, and it may not accurately reflect the actual visual quality and characteristics of the image.

Recently, Wang et al. [93] proposed an interesting UI measure based on the standard

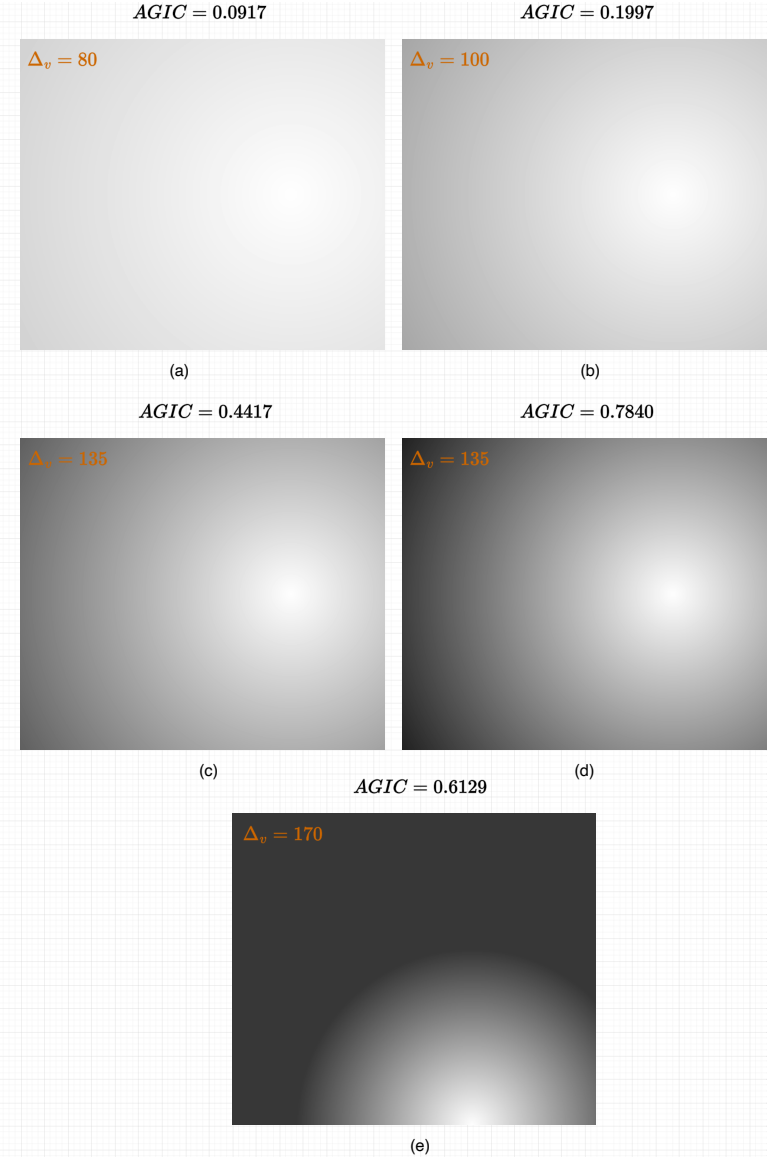


Figure 4.2: *The AGIC results on the synthesis image distorted by 4 levels of uneven illumination. Images (a) to (e) are the uneven illumination mask with a variant gradient level. From (a) to (d) the level of uneven illumination was increased one by one. The image (e) has the same gradient level compared to (d) but instead of using the same position, the mask is created with a different position to have a mask with a larger region of dark which is expected as an inefficient case of the AGIC. The corresponding AGIC indexes are at the top of each image using equation* (4.3).

deviation of the V component in the HSV (Hue-Saturation-Value) color space for microscopic images. They take the Standard Deviation of Value (SDV) channel of the HSV(hue-saturation-value) color space as an objective indicator to assess the image illumination. In HSV color space, V represents the brightness of pixel points. The SDV score is defined as follows:

$$SDV = \sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(V(i,j)-\mu_v)^2}{M \times N - 1}}, \tag{4.4}$$

where M, N represent the size of the V channel of the input image, μ_v represents its pixel mean, and $V(i,j)$ is the pixel value of i-th row and j-th column. It is capable of depicting the distribution of image illumination. The standard deviation reflects the degree of dispersion between the image pixel value and its pixel mean value. The smaller the standard deviation, the smaller the image contrast. Therefore, the smaller the SDV score, the more uniform the image illumination distribution. Fig. 4.3 shows us the used sample image in the SDV paper.

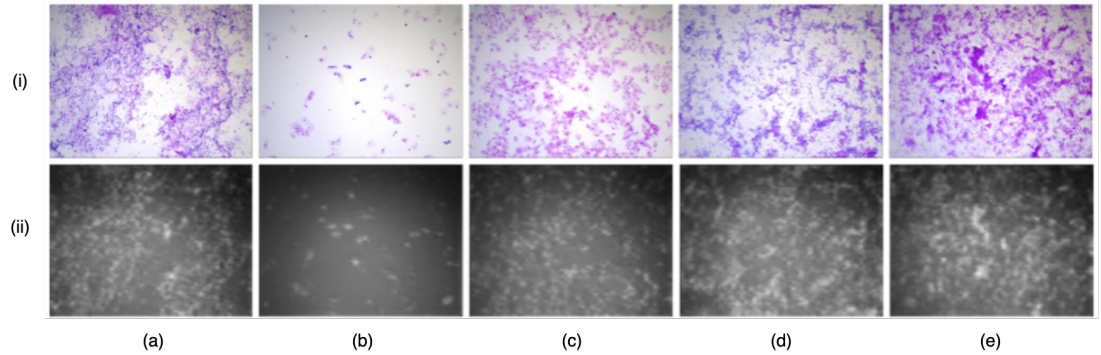


Figure 4.3: *Some image examples used in [93]. Row (i) is the original image. Row (ii) is the corresponding V-channel when they convert the image from RGB color space into HSV color space.*

Nonetheless, it's important to highlight that this method is directly applied to the V-channel of the image. Therefore, its effectiveness is more pronounced when dealing with images that possess a simple background and where the object within the image does not significantly impact the background illuminance (BI). However, when confronted with images featuring intricate backgrounds and multiple objects, the accuracy of this method can indeed diminish. To substantiate this point, the test results conducted on the LVQ dataset are presented in Figure 4.4. These results demonstrate how the method's performance may falter in scenarios characterized by complex backgrounds and numerous objects within the image.

As previously discussed, it's evident that the SDV method appears to be well-suited for images featuring uncomplicated backgrounds. However, in the context of WCE images, the presence of diverse backgrounds influenced by the surface and shape of the GI tract introduces a unique set of challenges. Directly applying the SDV method to the V-channel of such images can indeed pose difficulties when attempting to accurately estimate uneven illumination due to the intricate nature of these backgrounds.

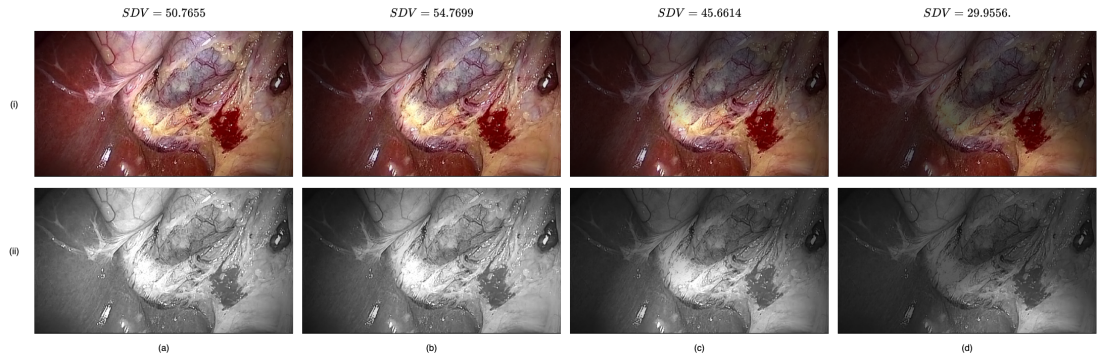


Figure 4.4: *The SDV result on the LVQ [257] dataset. Row (i) is the original image. Row (ii) is the corresponding V-channel when they convert the image from RGB color space into HSV color space. The corresponding SDV values are on the top of each column.*

All these methods and others have limitations both on the nature of the image signal and the field of application considered and more specifically the way the problem is formulated. Indeed, the few methods that seem interesting are based on optical models leading to an ill-posed problem that can only be solved by adding constraints or some priors [92] that cannot be rationally justified, or on data-dependence learning models [93].

In this work, we introduce a novel simple NR-IQA method based on the analysis of the local variations of BI components of images. The standard deviation of the difference signal between the BI values of the observed image and its histogram-equalized version is used as a physical measure to capture the effect of non-uniform lighting. The underlying idea is related to the observation that the more the image is affected by the UI effect, the more the contrast enhancement (CE) side-effect will accentuate the difference between the BI values of the image before and after CE, as illustrated in Fig. 4.6.

The main contributions of this work and the strengths of the proposed method are summarized and highlighted in the following.

- A new efficient NR-IQA metric for UI is proposed.
- The proposed metric outperforms the state-of-the-art methods when applied to UI in both laparoscopic images and WCE images.
- This metric could be adapted and applied to other imaging modalities.
- The proposed method does not require an image formation model, nor specific priors on the nature of the signal.

2 Proposed Method

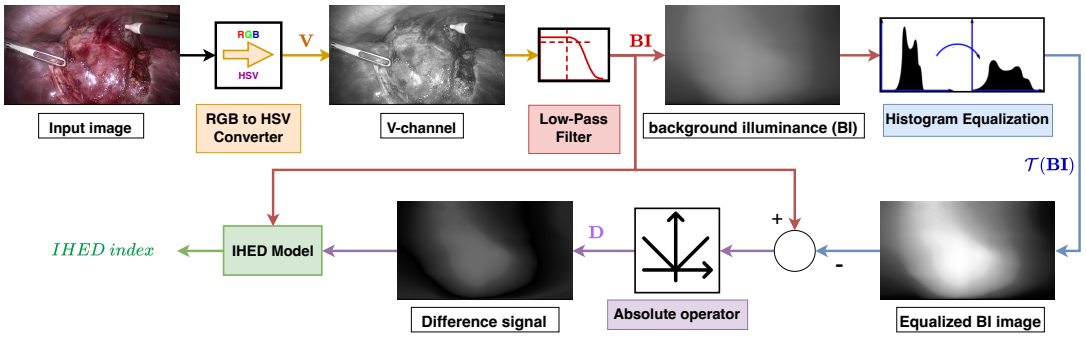


Figure 4.5: *Block diagram of the proposed uneven illumination assessment method.*

The proposed UI assessment metric is based on the analysis of the contrast enhancement effect on the degraded image. The basic idea is to quantify the effect of histogram equalization on the spatial distribution of BI. A similar idea has been proposed in a method for blind estimation of the blur in natural scenes [241].

Fig. 4.5 illustrates the process of uneven illumination assessment based on the analysis of the BI, estimated according to the algorithm described below.

i. The distorted image is converted to the HSV color space (Section 2.1).

ii. By assuming that the background illuminance is a slowly varying signal, as done in [258], the BI is estimated by filtering the image with a low-pass box filter of large size.

iii. Once the BI is estimated, the uneven illumination assessment metric is computed from the standard deviation of the difference between **BI** and its equalized version **E** (Section 2.3).

2.1 Color Space Conversion

As mentioned, the HSV color space is one of the most appropriate representations for quantifying color and luminance changes in an image [93]. Given a RGB color image of size $M \times N$, the three channels $\mathbf{R}$, $\mathbf{G}$ and $\mathbf{B}$ are first scaled into the range $[0,\ldots,1]$ resulting in:

$$[\mathbf{R}',\mathbf{G}',\mathbf{B}'] = [\frac{\mathbf{R}}{255},\frac{\mathbf{G}}{255},\frac{\mathbf{B}}{255}] \tag{4.5}$$

Then, the difference Δ between the largest and smallest scaled channel is extracted for each pixel:

$$\Delta = m_{max} - m_{min}, \tag{4.6}$$

$$m_{max} = \max(\mathbf{R}', \mathbf{G}', \mathbf{B}'), \tag{4.7}$$

$$m_{min} = \min(\mathbf{R}', \mathbf{G}', \mathbf{B}') \tag{4.8}$$

After that, the two smallest scaled channels are subtracted off, and divided by the difference between the largest and the smallest to generate the hue, $\mathbf{H} \in \mathbb{R}^{M \times N}$:

$$\mathbf{H}' = \begin{cases} undefined & if\ \Delta = 0 \\ \frac{\mathbf{G}' - \mathbf{B}'}{\Delta} & if\ m_{max} = \mathbf{R}' \\ \frac{\mathbf{B}' - \mathbf{R}'}{\Delta} + 2 & if\ m_{max} = \mathbf{G}' \\ \frac{\mathbf{R}' - \mathbf{G}'}{\Delta} + 4 & if\ m_{max} = \mathbf{B}' \end{cases} \tag{4.9}$$

$$\mathbf{H} = \mathbf{H}' \times scale_h, \tag{4.10}$$

The brightness, $\mathbf{V} \in \mathbb{R}^{M \times N}$, is based on the brightest colour channel:

$$\mathbf{V} = m_{max} \times scale_v \tag{4.11}$$

The saturation, $\mathbf{S} \in \mathbb{R}^{M \times N}$, is the difference between the largest and smallest colour channel values divided by the maximum channel value as:

$$\mathbf{S} = \begin{cases} 0 & m_{max} = 0 \\ \frac{\Delta}{m_{max}} \times scale_s & m_{max} \neq 0 \end{cases}, \tag{4.12}$$

where M, N is the width and height of the image and $scale_h$, $scale_s$, $scale_v$ represents the corresponding channel scale.

2.2 Background Illuminance Estimation

The estimation of background illuminance is a well-studied research topic in computer vision and several methods from the simplest to the most sophisticated ones have been proposed in the literature [259], [260]. The BI can be estimated using the variational framework for retinex (VFR) model [261] or empirical mode decomposition (EMD) [262]. However, these methods require solving a time-consuming optimization problem. We use a simple solution based on linear low-pass filtering (LPF). The background illuminance $\mathbf{BI} \in \mathbb{R}^{M \times N}$ is then estimated as follows:

$$\mathbf{BI} = G \circledast \mathbf{V}, \tag{4.13}$$

where $G \in \mathbb{R}^{l \times l}$ is the LPF kernel of size $l \times l$ and $\circledast$ denotes the convolution operator and $\mathbf{V}$ is the brightest component defined in (4.11). Here, a large-size kernel G is used to estimate the background illuminance:

$$l = \frac{1}{4}\min(M, N) \tag{4.14}$$

Fig. 4.6(d) - Fig. 4.6(f) illustrate some examples of the estimated BI maps.

2.3 Uneven Illumination Assessment

The first operation to be performed at this stage is the transformation of the input BI map by Histogram Equalization (HE).

Note that histogram equalization is a monotonic nonlinear pixel-value mapping aiming at transforming the pixel values of the input image to obtain an output image with uniform distribution. This operation has the effect of further accentuating the dark areas and brightening the light areas, thus generating an amplification of the global and local contrast.

Let $\mathcal{T}$ denote the histogram equalization transformation. The BI map is histogram equalized leading to $\mathbf{E} \in \mathbb{R}^{M \times N}$:

$$\mathbf{E}(i,j) = \mathcal{T}(\mathbf{BI}(i,j)) \tag{4.15}$$

Then, the difference signal $\mathbf{D} \in \mathbb{R}^{M \times N}$ is computed as:

$$\mathbf{D}(i,j) = \mid \mathbf{BI}(i,j) - \mathbf{E}(i,j) \mid \tag{4.16}$$

It can be observed through Fig. 4.6(j) - Fig. 4.6(l) that the proposed transformation has a different effect on the signal depending on the amount of UI on the input image.

As mentioned previously, the effect of HE is an amplification of the global and local contrasts as it gives more weight to bright pixels and low weight to dark pixels. As a consequence, it will highlight the UI effect as perceived through the equalized BI map. This observation inspired us to use the standard deviation, σ_D, of the difference signal $\mathbf{D}$, to quantify the UI effect.

$$\sigma_D = \sqrt{\frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(\mathbf{D}(i,j) - \mu_D)^2}{M \times N - 1}}, \tag{4.17}$$

where μ_D is the mean of $\mathbf{D}$. However, this index is inefficient in quantifying UI in the case of very low-light scenarios, when the background illuminance gradient is low and

makes the variations hardly noticeable. To address this issue, we introduce a novel index based on Illumination Histogram Equalization Difference (IHED). It is computed as follows:

$$IHED = \frac{\sqrt{\frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(\mathbf{D}(i,j)-\mu_D)^2}{M\times N-1}}}{\frac{1}{M\times N}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\mathbf{BI}(i,j)} \tag{4.18}$$

Note that (4.18) expresses the relationship between the uneven illumination effect and the equalized difference with regard to the BI map. High IHED values correspond to severe UI.

It should be noted that IHED was proposed before we established the QVCED database for WCE. Consequently, we began by assessing IHED using a dataset of laparoscopic images that was readily available. Subsequently, we extend its applicability and evaluate its effectiveness on the WCE dataset.

3 Performance Evaluation on laparoscopic images

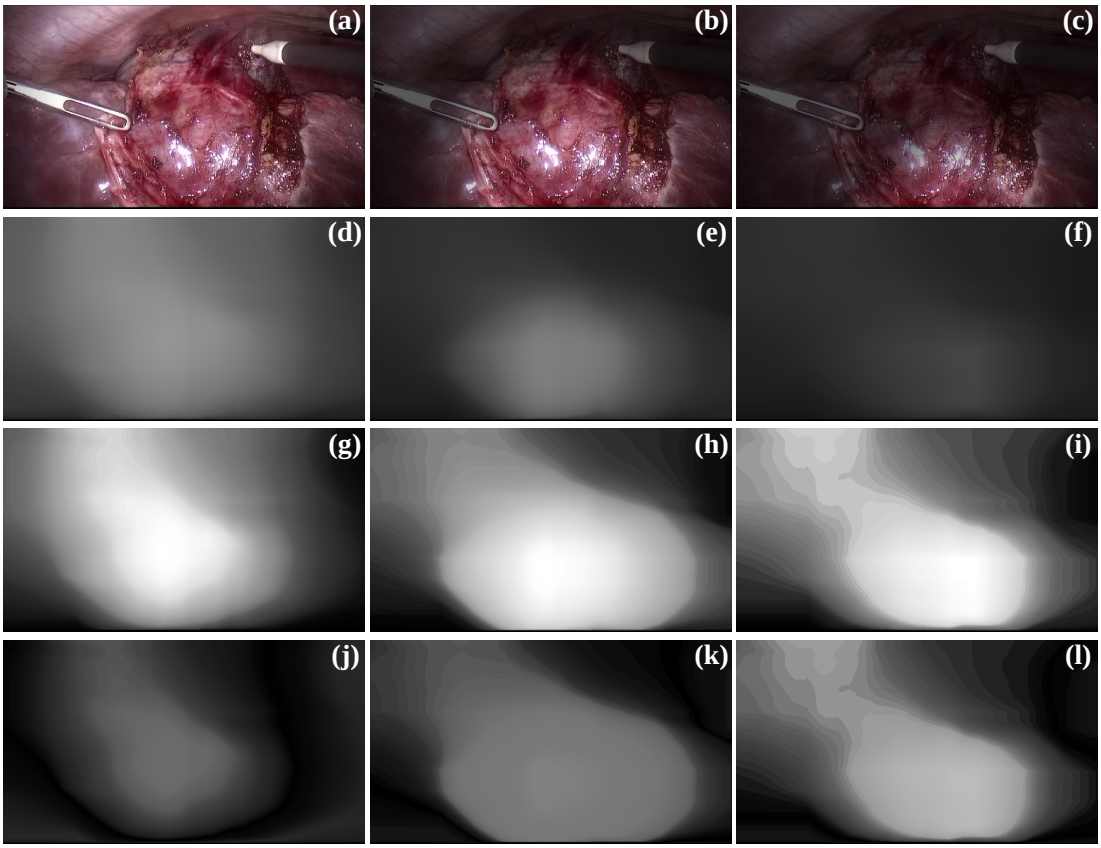


Figure 4.6: *Original images at different UI levels ((a)-(c)).The corresponding BI maps of* **V** *component ((d)-(f)). The equalized BI images (* **E** *signal) ((g)-(i)). The difference signals* **D** *between the BI maps (BI of the original and the corresponding equalized version) ((j)-(l)).*

In order to quantitatively evaluate the performance of the proposed method, firstly, the IHED is evaluated for assessing uneven illumination on laparoscopic images and is compared to both FR-IQA and NR-IQA methods.

3.1 Dataset

To quantitatively evaluate the performance of the proposed method on laparoscopic images, the LVQ [257] dataset is considered. This dataset contains 10 reference videos with 10 seconds duration and 25 fps frame rate. Each reference video is affected by UI at 4 different levels covering typical scenarios. The dataset contains the Mean Opinion Score (MOS) corresponding to 9 Expert observers' and 20 Non-Expert observers' evaluations. The subjective tests are performed in a controlled environment following the main ITU standard requirements. The details of the experimental setup are described in [257].

3.2 Correlation with Human Quality Judgment

As mentioned above, very few studies have been devoted to NR-IQA methods for UI. Therefore, we considered some non-specific distortions NR-IQA metrics including NIQE [28], VIIDEO [29], BRISQUE [73], and NIQE-K [263]. Besides these measures, two UI-dedicated metrics AGIC [92] and SDV [93] are also considered. To better evaluate the performance of the proposed index, we also use the conventional FR-IQA methods including PSNR and SSIM [26].

The consistency of the considered metrics is analyzed using two statistical criteria including Spearman's rank order correlation coefficient (SROCC), and Pearson's linear correlation coefficient (PLCC).

Table 4.1: *The correlation comparison on the LVQ dataset.*

		SROCC for expert scores	**SROCC for non-expert scores**	**PLCC for expert scores**	**PLCC for non-expert scores**
FR metrics	**PSNR**	0.9530	0.9372	0.9452	0.9561
	SSIM	**0.9580**	**0.9503**	**0.9847**	**0.9926**
NR metrics	**NIQE**	0.5620	0.5420	0.6655	0.6618
	NIQE-K	0.5828	0.5768	0.6936	0.6814
	BRISQUE	0.2635	0.2980	0.2973	0.3142
	VIIDEO	0.4281	0.3888	0.4035	0.3983
	AGIC	0.2191	0.1781	0.2231	0.2012
	SDV	0.6012	0.5477	0.6769	0.6311
	IHED (ours)	**0.9058**	**0.9047**	**0.8367**	**0.8571**

As can be seen from Table 4.1, IHED clearly outperforms the considered NR-IQA methods with a large margin and competes well with conventional FR-IQA metrics.

In order to further analyze the prediction performance of the proposed method, we provide a visual illustration through a scatter plot of the subjective rating (MOS) of both Expert and Non-Expert versus the objective score obtained with IHED on the LVQ dataset. As shown in Fig. 4.7, each point ('+') represents one test video. The red curve shown in Fig. 4.7 corresponds to the logistic function. IHED's points are close to each other and have a strong correlation with MOS. This further confirms the high performance of the proposed metric.

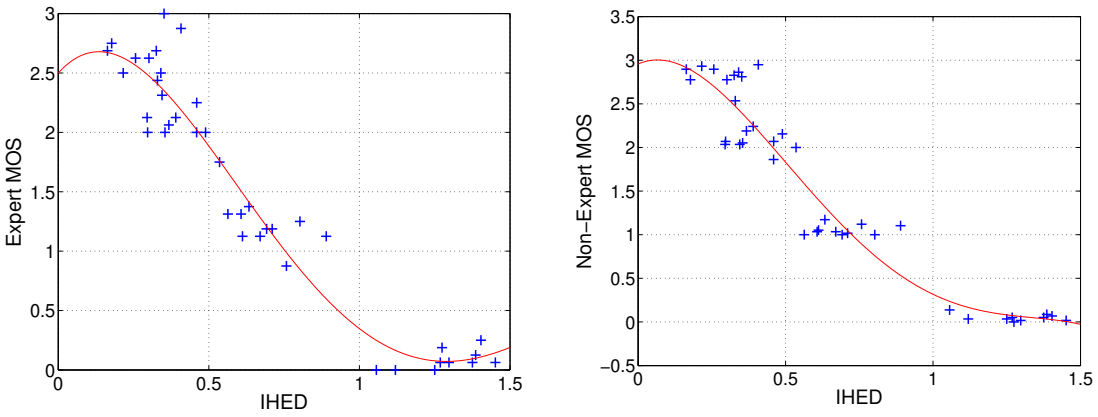


Figure 4.7: *Scatter plots of MOS versus prediction of IHED on the LVQ dataset. IHED versus the Expert MOS (left). IHED versus Non-Expert MOS (right).*

4 Performance Evaluation on WCE images

As mentioned in Chapter 3, the process of generating the distortion dataset involved simulating four types of distortion including noise, blur and uneven illumination, with each type having four levels of severity. This led to the inclusion of 80 distorted videos corresponding to uneven illumination within the QVCED dataset. The precise information such as the distortion types and their severity levels was presented in Table 3.1. A total of 34 participants, consisting of 12 experts and 22 non-experts, representing various age groups and backgrounds, participated in the subjective experiments.

4.1 Correlation with Human Quality Judgment

Firstly, the correlation with human quality judgment is also analyzed using two statistical measures including Spearman's rank order correlation coefficient (SROCC), and Pearson's linear correlation coefficient (PLCC) as shown in Table 4.2.

Table 4.2: *The correlation comparison on the QVCED dataset.*

		SROCC for expert scores	SROCC for non-expert scores	PLCC for expert scores	PLCC for non-expert scores
FR metrics	**PSNR**	0.9188	0.9202	0.9152	0.9116
	SSIM	**0.9322**	**0.9313**	**0.9471**	**0.9392**
NR metrics	**NIQE**	0.7621	0.7410	0.7555	0.7561
	NIQE-K	0.7448	0.7428	0.7336	0.7314
	BRISQUE	0.6692	0.6980	0.6837	0.6825
	VIIDEO	0.6681	0.6589	0.6521	0.6483
	AGIC	0.4277	0.4199	0.4281	0.4212
	SDV	0.8082	0.8117	0.8229	0.8311
	IHED (ours)	**0.9060**	**0.9061**	**0.8995**	**0.8970**

The IHED method also demonstrates its superiority when compared to other NR-IQA techniques due to the similarities between the laparoscopic and WCE image environments. The shared characteristics significantly contribute to the exceptional performance of IHED, making it a highly effective choice to evaluate the uneven illumination not only in both laparoscopic images and WCE images but also in other medical imaging modalities.

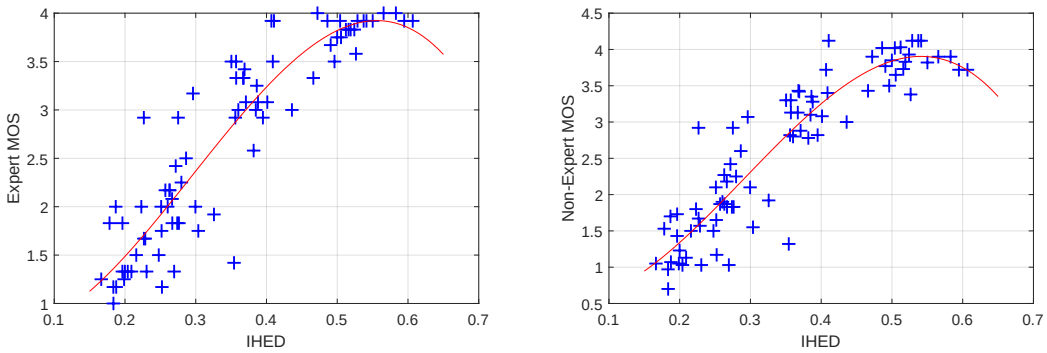


Figure 4.8: *Scatter plots of MOS versus prediction of IHED on the QVCED dataset. IHED versus the Expert MOS (left). IHED versus Non-Expert MOS (right).*

To further assess the predictive capabilities of our proposed method, we also present a visual representation using a scatter plot to compare the subjective ratings (MOS) from both Expert and Non-Expert participants to the objective scores obtained using IHED on the QVCED dataset. The red curve in Fig. 4.8 represents the logistic function in which IHED's data points are also closely clustered and exhibit a strong correlation with MOS.

5 Conclusion

This research endeavor has introduced a highly efficient No-Reference Image Quality Assessment (NR-IQA) method explicitly tailored to address the challenge of uneven illumination (UI) in laparoscopic images and WCE images. Our pioneering approach revolves around a meticulous examination of pixel value fluctuations within the background illuminance, meticulously extracted using histogram equalization. This innovative strategy leverages the inherent insensitivity of background illuminance to image content, effectively conquering the intricate challenges posed by complex backgrounds frequently encountered in the domain of medical imagery. Moreover, our novel metric incorporates the weighted average intensity of the background illuminance, equipping it with a significant advantage, especially in scenarios characterized by low-light conditions. It is noteworthy that our method not only outperforms existing NR-IQA techniques but also exhibits competitive performance comparable to FR-IQA methods. This underscores its potential for extensive application within the realm of medical image analysis, offering tangible benefits across a diverse spectrum of medical image types and scenarios. Consequently, there is a compelling case for the extension of this methodology to the assessment of image quality in other imaging modalities, where uneven illumination remains a pertinent concern.

Following the acquisition of distortion severity information, the subsequent chapter will exploit this data to introduce a novel image quality enhancement technique. This technique will be designed to effectively mitigate issues related to noise, blur, and uneven illumination, while accommodating variations in their levels across different scenarios. The proposed method will leverage the insights gained from the distortion severity assessment, allowing for tailored and adaptive image enhancement strategies that can significantly improve the quality of images captured in diverse conditions.

CHAPTER 5

The proposed image quality enhancement method: *TCFA: Triplet Clustering Fusion Autoencoder for Quality Enhancement of Wireless Capsule Endoscopy Images*

Abstract

Wireless Capsule Endoscopy (WCE) images are prone to degradations (such as noise, blur, or uneven illumination), which may adversely affect the reliability of the diagnosis. In this work, we propose a novel method, called **T**riplet **C**lustering **F**usion **A**utoencoder (TCFA), for enhancing the quality of WCE images. The method not only deals with the three types of degradations including noise, blur, and uneven illumination but also considers different levels of distortion severity. The main contributions presented in this work are the following: We introduce a distortion level encoder that classifies the severity of each type of distortion using the triplet loss function. Furthermore, we propose a variational cross-attention module that allows to extract both global and structural local information for efficient uneven illumination correction. Third, a pre-trained network is used to extract relevant structural features from WCE images. These contributions significantly improve the performance of image enhancement tasks, as demonstrated through extensive experiments on WCE images of the Kvasir-Capsule dataset. The proposed method is extensively evaluated on images with multilevel distortions, demonstrating superior performance compared to several state-of-the-art methods.[1]

[1] T. -S. Nguyen, M. Luong, J. Chaussard, A. Beghdadi, H. Zaag and T. Le-Tien *"TCFA: Triplet*

Chapter content

1 Introduction 92

2 Related Works 94

2.1 Denoising 95

2.2 Deblurring 95

2.3 Uneven Illumination Correction 96

3 Proposed Method 97

3.1 Overall Pipeline 97

3.2 Distorted Image Projector (DIP) 98

3.3 Distortion Level Encoder (DLE) 99

3.4 Variational Cross-Attention Module (VCAM) 99

3.5 Loss Function 100

4 Experiments 101

4.1 Dataset and Implementation Details 102

4.2 Ablation Study 103

4.2.1 Triplet Loss Margin 103

4.2.2 Effect of each TCFA's component 104

4.3 Comparison With State-of-the-Arts 106

4.3.1 Image Quality Assessment Comparison 106

4.3.2 Visual Quality Comparison 112

4.3.3 Statistical Comparison 112

5 Conclusion 118

Clustering Fusion Autoencoder for Quality Enhancement of Wireless Capsule Endoscopy Images," IEEE Transactions on Medical Imaging (Under preparation for submission).

1 Introduction

The introduction of Wireless Capsule Endoscopy (WCE) has completely transformed how we investigate and treat patients with suspected small bowel diseases. This technology, which first appeared in the year 2000, represents a significant advancement in small bowel examinations [230]. Unlike traditional endoscopy, which is limited to the upper GI tract (duodenum) and lower GI tract (terminal ileum), WCE allows us to see the entire gastrointestinal (GI) tract, including areas that were previously inaccessible [264]. Before the advent of WCE, the small intestine, which makes up 75% of the total length and 90% of the surface area of the GI tract, remained a challenging frontier for conventional endoscopy because it couldn't be directly visualized internally or in its entirety by any method [265].

While WCE has revolutionized and taken a prominent place in the screening and diagnosis of GI disease, a major challenge in WCE is its poor image quality, due to several physical factors depending on the acquisition system and the environment, which adversely affects diagnostic accuracy. Indeed, degradation is frequently produced during WCE image acquisition due to the physical limitations of the image sensors used and the uncontrolled acquisition environment. For example, WCE cameras with a narrow aperture and small sensors with a restricted dynamic range typically produce noise in the acquired frames [19]. Additionally, images captured under unstable circumstances exhibit excessive blurriness [231]. Significantly, the inherent motion of the capsule, caused by the peristaltic activity of the GI tract, combined with the light-coverage limitation of the capsule's camera [18] introduces uneven illumination. The existence of these distortions has compromised the quality and resulted in the efficiency of advanced activities, such as the identification, and monitoring of GI abnormalities. Therefore, a WCE image enhancement method is evidently required to improve the quality of the acquired images, further increasing the accuracy and reliability of GI lesion detection.

To mitigate the effect of distortions and improve the quality of WCE images, numerous algorithms have been proposed[266]–[268]. More precisely, Vani et al.[266] introduced a fusion technique that incorporates histogram equalization and color restoration, aiming to enhance contrast, reduce noise, and improve visibility in areas of lower illumination. Similarly, Long et al.[267] employed a fraction-power transformation incorporated with histogram modification, integrating a guide image, to increase contrast and improve image details in poor illumination scenarios. Besides, in [268], with the context of WCE videos, histogram mean and variance regarding each WCE frame are sequentially calculated to equalize the image intensity and smooth out hue fluctuations over time.

While the mentioned methods comparatively address general low-light situations, noise, or low contrast, they do present certain challenges. Specifically, gamma correction techniques can result in excessive amplification of pixel values and introduce various artifacts to the processed images. Additionally, when an image possesses a high dynamic range or includes regions with extremely high or low pixel values, histogram equalization can produce non-photorealistic areas and yield images that appear unnatural.

By considering the mentioned challenges, the recent advancements in deep learning, particularly in the field of deep convolutional neural networks (CNNs)[269], have placed a foundation for the development of new enhancement solutions. Current deep learning models, which can acquire generalizable priors from large-scale medical datasets, typically employ one among two architectures: An encoder-decoder or a generative model. The encoder-decoder models[270], [271] gradually map the input to an embedding latent space before gradually reverting to the original resolution. Although these methods can effectively reduce the dimensionality of input data, aiming to capture the most important features, the global contextual information is discarded, making image reconstruction exceedingly difficult. On the other side, the generative networks[152], [174] process images with more detailed spatial features. These networks, however, are less successful in the medical aspect since they generate uncontrolled features that result in less accurate outcomes in perspective medical tasks such as pathological classification. Recently, the attention-based methods[272], [273] were demonstrated to be effective in a variety of applications by exploring the positional information and the inter-pixel relationship.

Inspired by the attention-based advantages, in this work, we propose the TCFA method that addresses four levels of distortion severity and effectively handles different types of degradation, including noise, blur, and uneven illumination. It should be noted that, unlike the low-light condition where the entire image is dimly illuminated, uneven illumination presents a notable disparity. In the case of uneven illumination, areas in close proximity to the camera are excessively bright, while the distant regions are completely dark. Therefore, it is essential to take into account features relating to the precise region within the image affected by uneven illumination. This feature plays a critical role in developing an effective enhancement technique that specifically targets and addresses these affected regions.

In addition to the information about the affected regions, understanding the severity level of the distortions is equally significant. This information impacts the intensity of enhancement techniques applied. Different severity levels require varying degrees of correction methods to effectively mitigate the distortion effect presented in the image.

Therefore, considering both the affected regions and severity levels enhances the overall accuracy and efficacy of the enhancement process.

It should be noted that, in our case where we assume that uneven illumination forms a circular pattern, the distortion severity in each image patch is dependently related to the image patch position. Based on that idea, VCAM modifies the model's attention by aligning the level of uneven illumination for each image patch, represented as the attention key, based on the relationship to the patch's position, represented by the attention query, to bring the distribution of the attention key closer to a standard normal distribution. By imposing this constraint, the attention mechanism is guided to selectively focus on distorted regions in a more controlled and standardized manner. Additionally, we employ the perceptual loss function to secure the quality of enhanced images by considering dissimilarity in high-level features between the reconstructed images and the reference images [274]. Our primary contribution lies in an efficient enhancement method for improving WCE images. This method effectively manages diverse distortion levels through the application of the triplet loss function. Additionally, it introduces a variational cross-attention module that enhances cross-attention learning by constraining attention distribution. Furthermore, our approach incorporates a pre-trained network that preserves essential high-level image features.

The subsequent sections of this chapter are structured as follows. Section 2 presents the recent related enhancement methods applied on both WCE. Subsequently, Section 3 provides a detailed account of the proposed TCFA. Afterward, Section 4 focuses on the experimental results. In particular, Section 4.1 presents the implementation of the simulated dataset while Section 4.2 concentrates on the ablation study, illustrating the importance of each component of TCFA. The proposed method's performance is then presented and evaluated in Section 4.3. Finally, this chapter concludes in Section 5, summarizing the key findings and implications of this work.

2 Related Works

In this section, we present recent advancements in enhancement techniques employed in WCE. The focus of this section is to provide a comprehensive overview of the most relevant studies concerning denoising, deblurring, and addressing uneven illumination in WCE images.

2.1 Denoising

Noise is a commonly observed distortion that is frequently experienced in video acquisition and transmission systems and denoising has been an actively investigated subject within not only the medical image domain but also various image processing fields. State-of-the-art denoising approaches include conventional methods such as BM3D[275], total variation (TV)[276] and dictionary learning-based methods[277]. Although these methods have demonstrated promising results, the main difficulty is the ability to preserve subtle details. Recently, Convolutional Neural Networks (CNNs) have become popular for deep noise removal[142], [143] due to their ability to handle complex noise, but rely on high-quality noisy-clean image pairings for training, which is a non-trivial-problem. To better preserve fine spatial details, CNN-based multiscale techniques[144], [145] have been proposed. However, such models may fail in case of strong noise as they may be confused with the texture feature of the image.

Regarding WCE image denoising approaches, the most representative classical image processing methods applied to WCE images are the geometric mean filter[20], TV[110], and the wavelet-based approach[146]. More recently, the deep image prior (DIP) technique[19] followed by a blind image quality assessment network to iteratively generate the noisy image from the noise model and then a clear image from a noisy image. Nevertheless, the considered challenge is that numerous iterations are necessary to obtain an optimal reconstructed image from a noise model. Consequently, the aforementioned process is time-consuming and not efficient in practice.

2.2 Deblurring

Similar to denoising, much study has been done in recent decades with many different image deblurring solutions on both natural and medical imaging domains. Regarding natural images, image deblurring methods can be divided into classical image processing-based methods and CNN-based ones. A traditional method such as Total Variation (TV)[149], [155] involves using blind deconvolution techniques to estimate both the original image and the blur kernel concurrently. In the second category, as to CNN-based methods, Generative Adversarial Networks (GANs)[151], [152], have also been proposed to learn the end-to-end regression between a blurry input image and the corresponding sharp image. However, the generative model is not particularly effective in the medical field as it produces uncontrolled patterns that may pose a risk of inaccurate diagnosis. To overcome this challenge, Zhang et al.[153] demonstrated a hierarchical representation using a multi-patch network inspired by spatial pyramid matching to

handle blurry images on different scales. Lately,[154] incorporated deep U-net networks with the concept of self-attention technique to efficiently capture the local context of the degraded images.

Likewise, in the domain of WCE images, some efforts have been made for deblurring. Liu et al.[155] modified the updating iteration of the regular parameter in total variation using the fast iterative shrinkage/thresholding technique to deal with multi-channel images. In recent times, dictionary learning techniques[156], [157] have been employed to extract the inter-frame relation in WCE videos for super-resolution and synthesizing deblurring reconstructions.

2.3 Uneven Illumination Correction

Uneven illumination correction has long been a concern in medical image processing with a few classical methods such as filtering[165], [166], or retinex-based model[167]. Despite significant efforts towards this end, the number of approaches developed to address this challenge in WCE images remains limited. Long et al.[171] proposed to leverage the information from high-quality guided images of the same scene to improve the quality of low-quality WCE images by histogram comparison. The adaptive fraction-power transformation method modifies the power value based on local image features to achieve optimal enhancement results. However, the method's effectiveness may be limited in situations where guided images are not available or when significant variation exists between the guided images and the low-quality WCE images. In such cases, the approach may introduce artifacts or distortions into the enhanced images, thereby compromising their diagnostic accuracy. To address this challenge, Prasath et al.[16] proposed a modified feature-linking model (FLM)[172] that enhances image quality by converting RGB images to the HSV color space, thereby improving chromaticity and avoiding streaking artifacts commonly associated with histogram-based models. However, a limitation of FLM is its reliance on linking image features across multiple scales, which may not be effective in enhancing images with low contrast or complex background structures. Furthermore, the computational complexity of FLM may pose a limitation, particularly in real-time applications that require fast image processing.

In the closely related but not overlapping context, it should be noted that low light and uneven illumination are two distinct problems that can significantly degrade the image quality. They are independent problems that require different correction techniques to enhance the quality. In recent times, regarding the natural image domain, several low-light enhancement approaches[173]–[175] are also proposed to increase

visibility. In this work, our objective is also to evaluate the performance of these methods to determine their effectiveness in dealing with the problem of uneven illumination.

3 Proposed Method

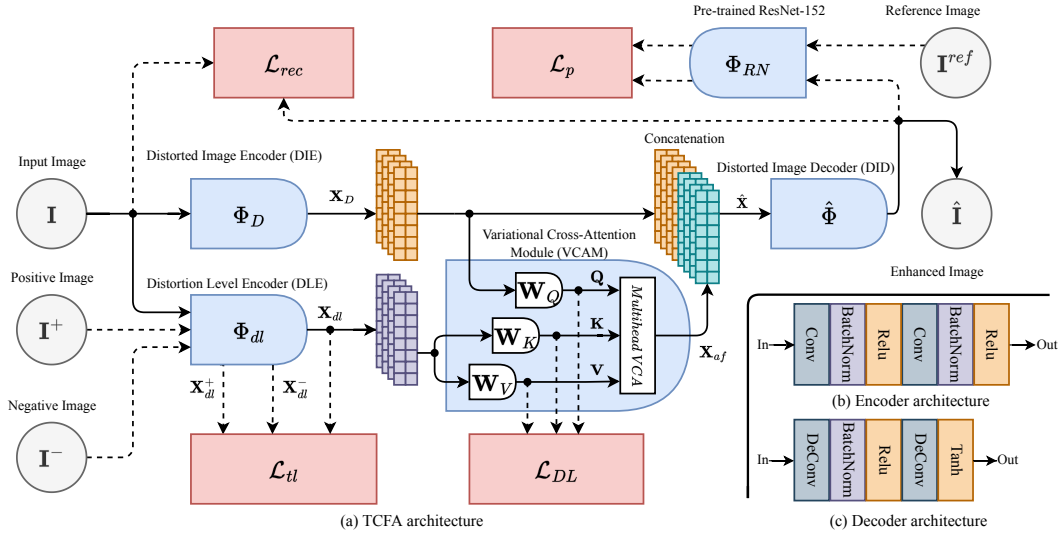


Figure 5.1: *(a) Overview of the proposed TCFA. Shaded circles represent processed images, bullet shapes represent deterministic functions, • represents a bifurcation, and red boxes represent cost terms. Dotted lines exclusively indicate the data path which is available only during the training phase. (b) The encoder architecture used in the DIE and DLE. (c) The employed decoder architecture used in the DID.*

In this section, we commence by presenting an overview of the proposed TCFA pipeline. Subsequently, we focus on each fundamental component comprising TCFA. then, we elucidate the formulation and analysis of the loss function in detail.

As depicted in Fig. 5.1 (a), TCFA encompasses a dual-branch autoencoder hierarchical network, encompassing three primary modules: the *Distorted Image Projector (DIP)* which consists of elements of the *Distorted Image Encoder (DIE)* and the *Distorted Image Decoder (DID)*; the *Distortion Level Encoder (DLE)*; and the *Variational Cross-Attention Module (VCAM)*.

3.1 Overall Pipeline

In this work, we present a novel two-branch method. The first branch prioritizes the preservation of local features through the implementation of the **Distorted Image Encoder (DIE)**. The local feature refers to specific characteristics or attributes within the image that are present on a smaller, localized scale within the image such as edges,

textures, or patterns. The second branch integrates the **Distortion Level Encoder (DLE)**, which captures and extract multi-level distortion feature by clustering the images into different groups related to distortion levels utilizing a triplet loss function (5.9), enabling effective handling of severity variations. Subsequently, the two features are inputs into the innovative **Variational Cross-Attention Module (VCAM)**. This module, inspired by the cross-attention technique[278], considers both the local image features (obtained from DIE) and distortion severity information (obtained from DLE), thereby leading to improved enhancement outcomes. Finally, the **Distorted Image Encoder (DID)** is utilized to decode the concatenation of local image features (obtained from DIE) and attention-fused features (obtained from VCAM) to reconstruct the enhanced image.

Given a degraded image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, the DIE Φ_D extracts the local features $\mathbf{X}_D \in \mathbb{R}^{C\times H_x\times W_x}$ in the latent space. $C = 64, H_x = W_x = \frac{H}{4}$ are feature dimensions, $H = W = 336$ are the height and width of the input image. Simultaneously, the DLE Φ_{dl} is applied to $\mathbf{I}$ to extract the distortion level information $\mathbf{X}_{dl} \in \mathbb{R}^{C\times H_x\times W_x}$. Next, the VCAM utilizes a cross-attention mechanism to capture positional dependencies between the distortion severity and the corresponding image patch position to generate attention-fused features $\mathbf{X}_{af} \in \mathbb{R}^{C\times H_x\times W_x}$. Finally, the DID $\hat{\Phi}$ is utilized to decode the concatenation $\hat{\mathbf{X}} \in \mathbb{R}^{2C\times H_x\times W_x}$, $\hat{\mathbf{X}} = concat(\mathbf{X}_D, \mathbf{X}_{af})$ of distorted image features and attention-fused features to obtain the enhanced image $\hat{\mathbf{I}}$.

3.2 Distorted Image Projector (DIP)

As mentioned before, the proposed DIP includes the Distorted Image Encoder (DIE) using the CNN backbone, as shown in Fig. 5.1(b), which contains two convolution blocks. Each convolution block consists of one convolution layer followed by Batch Normalization layers and a Rectified Linear Unit (RELU) activation function. The DIE module typically converts the input image $\mathbf{I}$ into the latent space, thus extracting local features $\mathbf{X}_D$.

On the other hand, the Distorted Image Decoder (DID) is implemented to reconstruct the enhanced images using the decoder architecture (Fig. 5.1(c)) of two deconvolution layers or transposed convolution layers. It upsamples the concatenated features $\hat{\mathbf{X}} \in \mathbb{R}^{2C\times H_x\times W_x}$, where feature dimension C is set to 64 for the first deconvolution layer and 3 in the following deconvolution layer, to generate the output image with the same dimensions as the input. The kernel size is set as 3, stride as 2, and the momentum used in Batch Normalization is set to 0.1 for all convolution and deconvolution layers.

3.3 Distortion Level Encoder (DLE)

We propose the DLE to encode the distorted image into another latent space where the triplet loss function[279] clusters the distortion level features into different populations. Similar to DIE, the DLE consists of two convolution blocks, and each block contains a convolution layer followed by a Batch Normalization layer with a RELU activation function (with $C = 64, k = 3, s = 2$).

More precisely, the DLE is trained using the triplet loss which encourages that dissimilar pairs be distant from any similar pairs by at least a certain margin value as:

$$l_{triplet} = \max\Big(\Big\|\mathbf{X}_{dl} - \mathbf{X}_{dl}^{+}\Big\|_F^2 - \Big\|\mathbf{X}_{dl} - \mathbf{X}_{dl}^{-}\Big\|_F^2 + m, 0\Big), \tag{5.1}$$

where $\mathbf{I}$ is the input image, and $\mathbf{X}_{dl} \in \mathbb{R}^{C\times H\times W}$ is the extracted distortion level feature of $\mathbf{I}$. The positive image $\mathbf{I}^{+}$ is a randomly chosen image by the DLE Φ_{dl}, having the same distortion level as $\mathbf{I}$, and $\mathbf{X}_{dl}^{+}$ is the extracted distortion level feature of $\mathbf{I}^{+}$. On the other hand, the negative image $\mathbf{I}^{-}$ is a randomly-chosen image by the DLE Φ_{dl}, having a different distortion level from $\mathbf{I}$, with m as the margin hyperparameter, and $\mathbf{X}_{dl}^{-}$ is the extracted distortion level feature of $\mathbf{I}^{-}$.

3.4 Variational Cross-Attention Module (VCAM)

Besides noise and blur, we also target uneven illumination correction. Uneven illumination correction is challenging due to the requirement for global information from the entire image. It is necessary to utilize both the distorted image feature $\mathbf{X}_D$ and the distortion level feature $\mathbf{X}_{dl}$. Inspired by[278], we designed the VCAM to establish a correspondence between them and perform feature aggregation.

First, the distorted image feature $\mathbf{X}_D \in \mathbb{R}^{C\times H_x\times W_x}$ and the distortion level feature $\mathbf{X}_{dl} \in \mathbb{R}^{C\times H_x\times W_x}$ are divided into non-overlapping local windows of size $M \times M$ with a total number of $\frac{H_x W_x}{M^2}$ windows. In this work, $W_x = H_x = \frac{H}{4}$ and $M = 8$. Then, the attention mechanism jointly learns the $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d_{model}}, d = M^2C, \mathbf{W}_P \in \mathbb{R}^{\frac{H_x W_x}{M^2}\times d_{model}}$ which are the query (Q), key (K), value (V) and position weight matrices with $d_{model} = 512$, respectively.

$$\begin{aligned}\mathbf{Q} &= \mathbf{X}_D\mathbf{W}_Q + \mathbf{W}_P,\\ \mathbf{K} &= \mathbf{X}_{dl}\mathbf{W}_K + \mathbf{W}_P,\\ \mathbf{V} &= \mathbf{X}_{dl}\mathbf{W}_V + \mathbf{W}_P\end{aligned} \tag{5.2}$$

Additionally, in this work when the uneven illumination is supposed to have a circular-shape background illuminance as appears in (5.14), it is noticeable that the level of distortion present in each image patch has a dependent relationship with its position within the image. To guide the network to focus on the distorted regions in a controlled manner by capturing that relationship, we apply a parameter transformation that enforces the distribution of the key generated from the distortion level feature to be close to a standard normal distribution:

$$\mathbf{K}' = \mathbf{K} + \mathbf{Q} * \epsilon, \tag{5.3}$$

where $\epsilon \sim \mathcal{N}(0,1)$.

Subsequently, the multi-head variational cross-attention module (Multihead VCAM) is used to extract the multi-head attention information $\mathbf{Z} \in \mathbb{R}^{C \times d}$, from various representation subspaces, which is divided into h heads for learning in parallel (in this work, $h = 8$), and each head's output dimension is $d_{head} = \frac{d_{model}}{h}$.

$$\mathbf{Z} = concat(head_1, \ldots, head_h)\mathbf{W}^O \tag{5.4}$$

$$head_i = VarCrossAtten(\mathbf{W}_i^Q\mathbf{Q}, W_i^K\mathbf{K}', W_i^V\mathbf{V}), \tag{5.5}$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d_{model} \times d_{head}}$, $\mathbf{W}^O \in \mathbb{R}^{d_{model} \times d}$ are learnable weights. The attention is computed by encoding the distorted image feature as queries and the distortion level feature as both keys and values:

$$VarCrossAtten(\mathbf{Q}, \mathbf{K}', \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}'^T}{\sqrt{d_{head}}}\right)\mathbf{V} \tag{5.6}$$

Afterward, the LayerNorm (LN) layer is added and the residual connection is employed. Finally, the output attention-fused feature is defined as:

$$\mathbf{X}_{af} = LN(\mathbf{Z} + \mathbf{X}_D) \odot \mathbf{X}_{dl}, \tag{5.7}$$

where $\odot$ denotes the element-wise multiplication.

3.5 Loss Function

The proposed TCFA model will be trained using an appropriate loss function, with given input distorted images $\mathbf{I}_i \in \mathbb{R}^{3 \times H \times W}, i = 1 \ldots N$, N is the number of distorted images, which aims to achieve the following objectives:

- Learning a useful representation $\hat{\mathbf{I}} \in \mathbb{R}^{3 \times H \times W}$ of the distorted input image through a nonlinear dimensionality reduction by autoencoder (AE). Hence, the loss related

to the reconstruction error of the AE, $\mathcal{L}_{rec}$, is given by the dissimilarity between distorted and enhanced images.

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{i=1}^{N}\left(\left\|\mathbf{I}_i - \hat{\mathbf{I}}_i\right\|_F^2\right) \tag{5.8}$$

- Minimizing the distance between extracted distortion features of the random similar image pairs $(\mathbf{I}_i, \mathbf{I}_\mathbf{i}^+)$ having similar distortion level and maximizing the distance between extracted distortion features of the random dissimilar image pairs $(\mathbf{I}_i, \mathbf{I}_\mathbf{i}^-)$ having different distortion level by Distorted Level Encoder $\mathbf{\Phi}_{dl}$ and at least a certain margin value m.

$$\begin{aligned}\mathcal{L}_{tl} = \sum_{i=1}^{N}\max\Big(&\left\|\Phi_{dl}(\mathbf{I}_i) - \Phi_{dl}(\mathbf{I}_i^+)\right\|_F^2 \\ &- \left\|\Phi_{dl}(\mathbf{I}_i) - \Phi_{dl}(\mathbf{I}_i^-)\right\|_F^2 + m, 0\Big)\end{aligned} \tag{5.9}$$

- Estimating the difference between the generated attention key distribution and the standard normal distribution. To estimate it, the Kullback-Leibler divergence loss $\mathcal{L}_{DL}$ is considered as the dissimilarity Euclidean distance:

$$\mathcal{L}_{DL} = -\frac{1}{2}\sum_{i=1}^{N}\left(1 + log(\|\mathbf{Q}_i\|_F^2) - \|\mathbf{K}_i\|_F^2 - \|\mathbf{Q}_i\|_F^2\right), \tag{5.10}$$

  where $\mathbf{Q}_i$ and $\mathbf{K}_i$ are the query and key generated by the VCAM corresponding to the image $\mathbf{I}_i$, respectively.

- Learning the high-level image feature representations extracted from pre-trained convolutional neural networks which are essential for further image reconstruction. The perceptual loss $\mathcal{L}_p$ is defined as the dissimilarity between the features of the enhanced images $\mathbf{I}_i^{ref}$ and reference images generated by the ResNet-152 network Φ_{RN} pre-trained on the ImageNet dataset:

$$\mathcal{L}_p = \frac{1}{N}\sum_{i=1}^{N}\left(\left\|\Phi_{RN}(\mathbf{I}_i) - \Phi_{RN}(\mathbf{I}_i^{ref})\right\|_F^2\right) \tag{5.11}$$

  The global loss function is defined as:

$$\mathcal{L}_t = \mathcal{L}_{rec} + \mathcal{L}_{tl} + \mathcal{L}_{DL} + \mathcal{L}_p \tag{5.12}$$

4 Experiments

In this section, we first provide details of the experimental setup including the dataset and environment which are employed in our study. Subsequently, we conduct a series of

comprehensive ablation studies to thoroughly evaluate each component of our proposed TCFA. Finally, we proceed to assess the effectiveness and efficiency of TCFA by applying it to various image enhancement tasks using simulated datasets. A comparative analysis is then conducted to assess its performance compared to recent works in the considered field.

4.1 Dataset and Implementation Details

To create the environment, Adam optimizer[280] is used for training and the model uses a batch size of 32 and a learning rate of 10^{-3} for 100 epochs. For a fair comparison, the competing methods all use their default parameter settings. Code will be available at https://github.com/tansyab1/TCFA.

For the dataset, we evaluate the proposed method on the proposed distortion dataset QVCED (Chapter 3). In this work, in each type of distortion including Additive White Gaussian Noise (AWGN), Defocus Blur, and Uneven Illumination, four severity levels are considered and applied to generate the distorted images. For each level of distortions, the training, validation, and testing set contain respectively 5000, 500, and 4000 images. These selected images are classified as having the best quality in QVCED to be used as reference images for distortion simulation. After preparing the reference images, the distortion is then added to generate the dataset that will be used in the subsequent experiments.

To simulate the noise, the Additive White Gaussian Noise (AWGN) model is incorporated into our dataset which assumes that the noise present in the video follows a Gaussian distribution. To control the severity, AWGN level is configured with $\sigma_n \in \{5, 10, 20, 30\}$.

Regarding the defocus blur, a commonly used approach involves incorporating an isotropic Gaussian kernel to accurately model the rotational symmetry around the optical axis that arises from the blurring effect caused by simulated defocusing of the lens. The corresponding impulse response, denoted as $h_{db}(x, y)$, can be mathematically expressed as follows:

$$h_{db}(x,y) = \frac{1}{2\pi\sigma_{db}^2} \exp \frac{-(x^2+y^2)}{2\sigma_{db}^2}, \ \sigma_{db} \in \{1,2,3,5\} \tag{5.13}$$

To simulate the uneven illumination, the reference image is first converted from RGB color space to HSV color space. A normalized circular gradient mask is then applied to the V-channel. The simulated circle-gradient mask $\mathbf{M}(i,j) \in \mathbb{R}^{H \times W}$ that matches the

dimensions of the original image, is defined by:

$$\mathbf{M}(i,j) = 255 - \left[\frac{2\Delta_I}{W}\sqrt{(i-i_c)^2+(j-j_c)^2}\right], \tag{5.14}$$

where $\mathbf{M}(i_c, j_c)$ is the circle center in $(i_c, j_c) \in \{(112,224),(168,168),(224,224)\}$. Different levels of illumination are achieved by controlling the difference of intensity between the central pixel of the image and the darkest pixel of the border and setting it in $\Delta_I \in \{80,100,135,170\}$.

Once the environment and the dataset are prepared, we proceed to conduct an ablation study aimed at analyzing the hyperparameters of the loss function and evaluating the significance of each component within the TCFA scheme. The objective of this study is to gain insights into the impact of different hyperparameter settings on the overall performance of TCFA, as well as to assess the individual contributions of each component in achieving the desired image enhancement outcomes.

4.2 Ablation Study

Within this subsection, our analysis focuses on examining the influence of the margin parameter on the triplet loss function in order to determine the margin that generates the distortion level feature with the most effective discrimination. Moreover, we thoroughly scrutinize each individual component of the TCFA, providing a comprehensive and detailed analysis. Furthermore, our evaluations encompass image denoising, deblurring, and the correction of uneven illumination, utilizing a range of diverse quality assessment metrics.

4.2.1 Triplet Loss Margin

By employing the proposed Distortion Level Encoder, which follows the triplet loss function defined in (5.9), the margin parameter denoted as m assumes a crucial role and necessitates meticulous selection. The reason behind this importance lies in its potential impact on the discrimination capabilities of the extracted features. Consequently, a comprehensive evaluation is conducted through a consequent test, wherein the margin parameter $m \in \{1,2,\ldots,32,64\}$ is varied to determine the optimal margin value that produces the most favorable outcomes in terms of discrimination. To assess the discrimination, evaluation metrics such as the Davies-Bouldin score[281] and Silhouette score[282] are employed where a lower Davies-Bouldin score and higher Silhouette score indicate better clustering. By utilizing these evaluation metrics, this study can quantitatively assess the quality of the extracted features and determine the most suitable margin for each type of distortion data. The obtained comparison results are

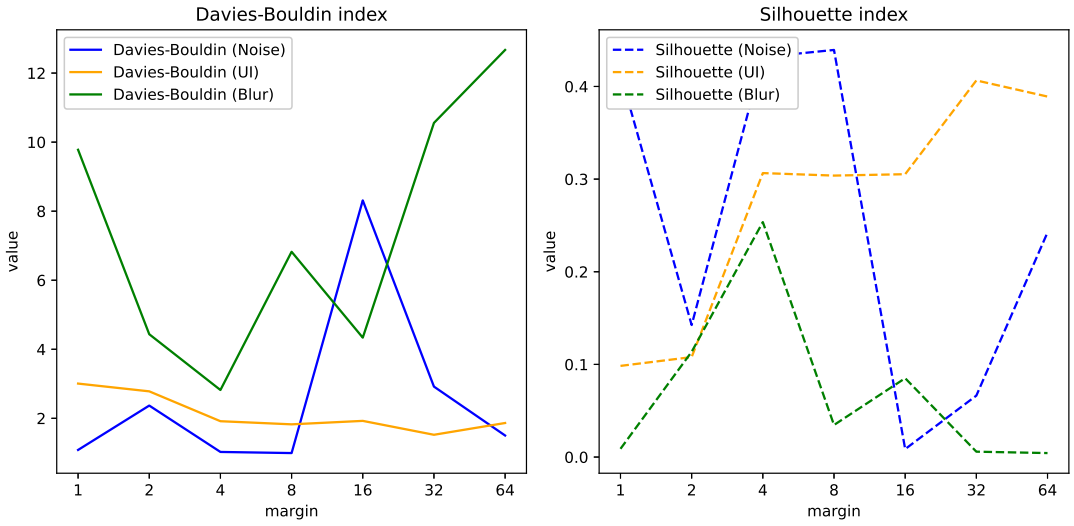


Figure 5.2: *Quantifying discrimination in clustering results using Davies-Bouldin score and Silhouette score.*

presented in Fig. 5.2. Notably, it is noticeable that the margin value of $m = \{8, 4, 32\}$ exhibits the best discrimination for the triplet loss regarding noise, blur, and uneven illumination, respectively. Additionally, a decrease in discrimination performance is observed after increasing the margin after the selected value. This occurrence indicates that setting the margin too high, beyond the discriminative capacity of the Distortion Level Encoder, adversely affects the performance. Consequently, this study emphasizes the importance of selecting appropriate margin values while considering the trade-off between discrimination ability and architecture. Based on these findings, the consequent experiment of each kind of distortion will be conducted using the corresponding margin value.

4.2.2 Effect of each TCFA's component

In this study, we would like to analyze the effect of individual sub-components of TCFA, specifically the VCAM, and DLE, on the performance of the TCFA model. Firstly, we devised an architecture without attention mechanisms, incorporating only the DIP and DLE. The features extracted from the DLE are directly combined with one of the DIE. Subsequently, the integrated features are then fed into the DID for further processing. This architecture allows focusing on the importance of the VCAM while leveraging the capabilities of the DIP and DLE for feature extraction and representation.

Secondly, as part of our modifications, we have similarly decided to remove the DLE from the architecture. As a result, the architecture now primarily focuses on the

remaining components, named self-attention architecture, wherein the features extracted solely from the DIE are directly supplied to the VCAM. By implementing this revised architecture, our objective is to thoroughly investigate and assess the significance and impact of the DLE in contributing valuable features to the attention-based architecture.

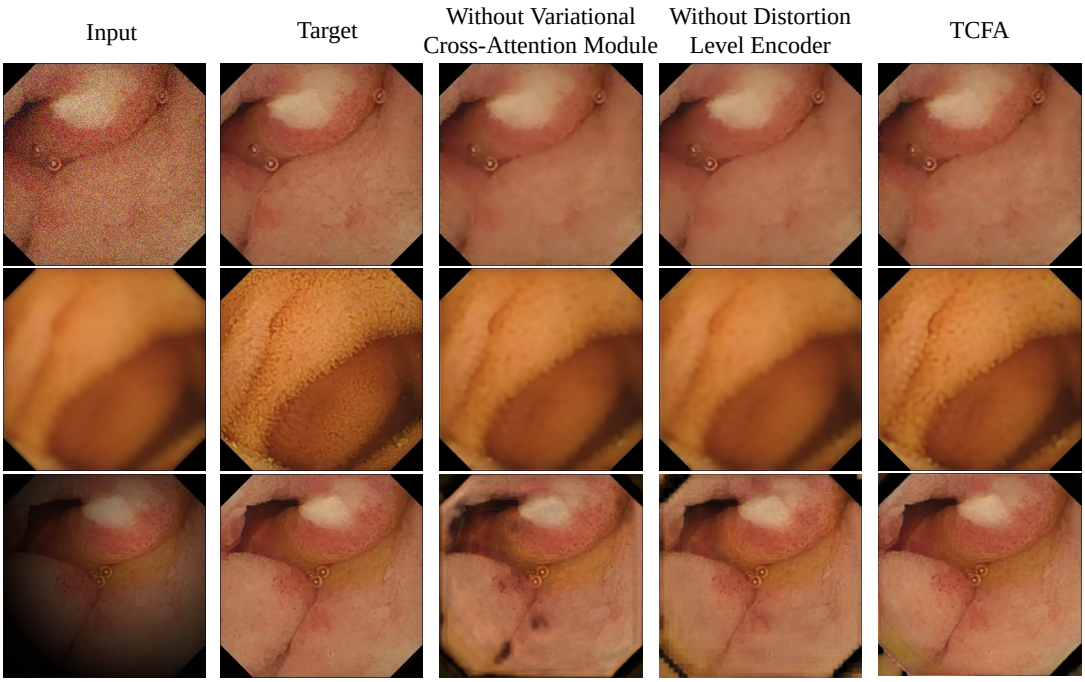


Figure 5.3: *Visual comparisons obtained from the ablation study conducted on TCFA. Rows 1 to 3 represent different elements, including the distorted image inputs, the reference image, outputs from TCFA without the VCAM, outputs from TCFA without the Distortion Level Encoder, and outputs from the final version of TCFA, respectively. The visual assessment focuses on addressing issues such as noise, blur, and uneven illumination. Notably, the final version of TCFA demonstrates the most significant improvement in mitigating these aforementioned issues and produces visually appealing results. For a closer examination of the image details, please zoom in.*

As shown in Fig.5.3, the outcomes obtained from both the non-attention and self-attention architectures are unsatisfactory, highlighting the effectiveness of TCFA in image enhancement, particularly in the task of correcting uneven illumination. Unlike noise and blur, where either global contextual features or local features alone may not be enough, both global contextual features and local features must be taken into account. The local feature extracted through the convolution operation captures the relationship between neighboring regions, highlighting their smooth intensity transition. On the other hand, the global feature focuses on enhancing the overall quality of specific regions. Therefore, the absence of either feature will lead to an unsatisfactory outcome, especially when dealing with uneven illumination.

Indeed, the absence of the VCAM (as seen in the third column) leads to images with black regions, indicating that the enhancer only relies on local features from a local image patch. In contrast, the attention-based architecture without the DLE focuses on global structural features and improves results compared to the non-attention architecture (fourth column). However, the outcome is still imperfect due to the lack of smooth transitions between regions which generates the box effect easily to be observed in the outer of the image, which arises from the absence of local features. Evidently, these results indicate that introducing distortion-level information into the cross-attention mechanism is beneficial for image enhancement tasks. This suggests that immigrating the knowledge of distortion levels into the attention mechanism can improve the network's focus in a controlled manner compared to original self-attention, and then effectively address WCE image distortion associated with different severity.

4.3 Comparison With State-of-the-Arts

This subsection involves an evaluation of TCFA's performance in comparison to other advanced methods. We carried out various experiments to achieve this, such as analyzing visual quality, statistical comparison, and performing full-reference/no-referenced image quality assessment comparisons. The following sections will provide more details on these experiments. More precisely, we comprehensively compare the proposed method against a range of denoising, blurring, and uneven illumination correction techniques. Additionally, we assess the performance of several low-light enhancement methods to determine their efficacy in addressing the uneven illumination problem. By emphasizing the disparity between low-light conditions and uneven illumination, we can obtain a more thorough understanding of the relative strengths and limitations of each kind of method, as well as its potential practical applications.

4.3.1 Image Quality Assessment Comparison

We compare our method with recent state-of-the-art methods which focus on denoising, deblurring, and uneven illumination correction. Furthermore, we aim to validate the effectiveness of the low-light enhancement methods in scenarios with uneven illumination. First, three full-reference metrics that are considered in the comparison experiments are Peak signal-to-noise ratio (PSNR), the structural similarity index measure (SSIM), and Visual information fidelity (VIF). Second, in situations where it is not possible to access the original reference content or when there are limitations in terms of storing or transmitting reference content, the use of no-reference quality metrics becomes especially significant. Therefore, to evaluate the performance of the proposed method in

the enhancement task, two No-Reference metrics including Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), and Natural Image Quality Evaluator (NIQE) are also considered. It should be noted that the high value of NIQE presents worse quality. Moreover, to illustrate the superiority of the proposed method on specific uneven illumination problems, the metric known as LOE (Lightness Order Error) is taken into account to assess the enhanced image quality. LOE plays a crucial role in measuring the degree of distortion in the lightness of the enhanced images, with lower LOE values indicating superior enhancement results.

In addition, computing the number of images with the best quality can provide valuable insights into the overall performance of the enhancement algorithm. By analyzing the method having the highest number of enhanced images that achieve the best quality, we can evaluate the effectiveness of the algorithm in improving image quality and determine whether it is suitable for practical applications. As a result, we will incorporate this quantity test (QT) of analyzing the number of images with the best quality into our assessment of various no-reference metrics (as shown in Fig. 5.4 to Fig. 5.6. This will enable us to obtain a comprehensive understanding of the algorithm's performance and make well-informed decisions regarding its practical application.

Table 5.1: *Quantitative comparison with state-of-the-art denoising methods in terms of five image quality assessment metrics*

	Image Quality Assessment Metrics				
	PSNR↑	SSIM↑	VIF↑	BRISQUE↑	NIQE↓
BM3D[275]	35.33	0.9079	0.6818	56.09	9.99
CycleISP[142]	33.65	0.9164	0.6429	51.27	9.97
DANet[143]	35.79	0.8864	0.6291	39.29	6.95
DIPNet[19]	37.43	0.9245	0.6010	55.73	13.26
VDNet[283]	36.83	0.9344	0.6496	49.79	13.10
MPRNet[284]	36.34	0.8889	0.6260	39.66	8.65
MIRNet[144]	39.01	0.9364	0.6785	51.65	8.59
Uformer[154]	38.43	0.9483	0.6907	52.19	6.83
TCFA	**39.34**	**0.9761**	**0.7055**	**56.25**	**6.37**

Our proposed TCFA method outperforms the compared reconstruction methods in all metrics and quantity tests also, as shown in Table 5.1 to 5.3, demonstrating its

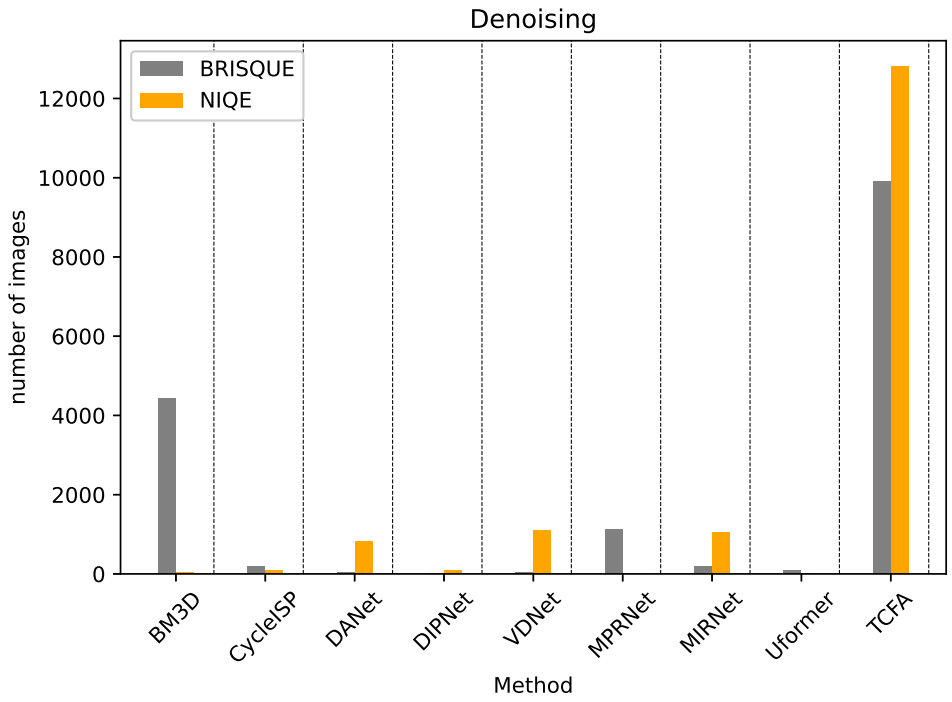


Figure 5.4: *Quantitative comparison with state-of-the-art denoising methods in terms of the quantity test on BRISQUE/NIQE.*

Table 5.2: *Quantitative comparison with state-of-the-art deblurring methods in terms of five image quality assessment metrics.*

	Image Quality Assessment Metrics				
	PSNR↑	SSIM↑	VIF↑	BRISQUE↑	NIQE↓
TV[155]	31.40	0.9108	0.6575	49.40	10.77
DBGAN[151]	37.75	0.9271	0.6350	54.72	10.27
Deblur-GANv2[285]	37.43	0.9156	0.6282	58.01	10.50
DMPHN[153]	37.92	0.9346	0.6465	54.85	10.26
MPRNet[284]	38.04	0.9271	0.6430	57.43	10.49
MIRNet[144]	**40.07**	0.9259	0.7074	59.38	9.05
Uformer[154]	39.49	0.9433	0.7457	59.36	9.21
TCFA	40.05	**0.9518**	**0.7517**	**60.06**	**8.61**

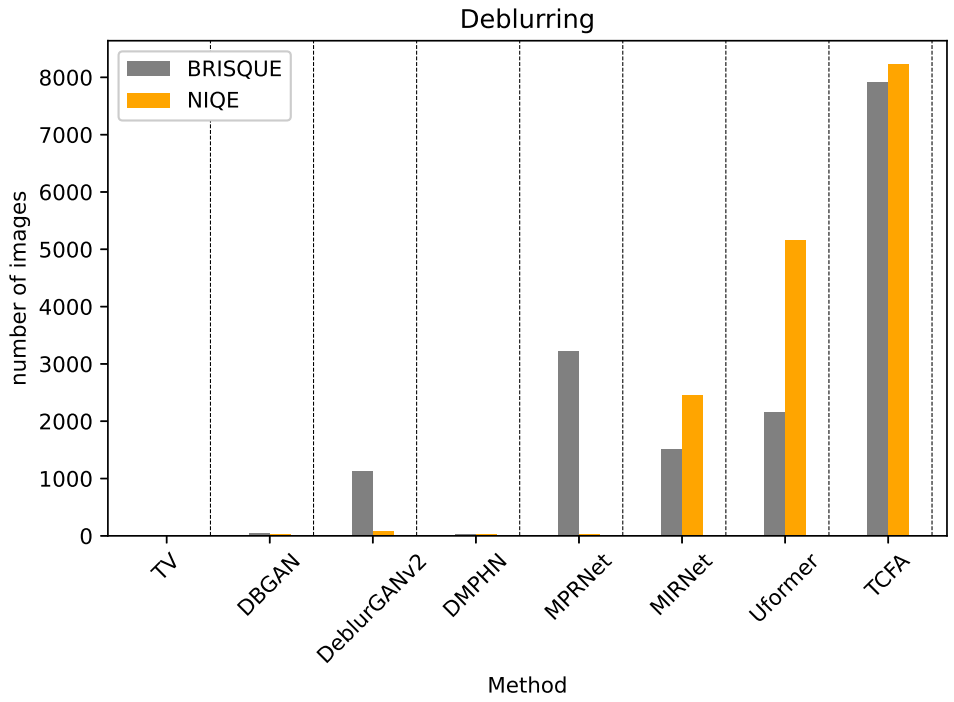


Figure 5.5: *Quantitative comparison with state-of-the-art deblurring methods in terms of the quantity test on BRISQUE/NIQE.*

Table 5.3: *Quantitative comparison with state-of-the-art uneven illumination correction methods in terms of six image quality assessment metrics.*

	Image Quality Assessment Metrics					
	PSNR↑	SSIM↑	VIF↑	BRISQUE↑	NIQE↓	LOE↓
AFGT[171]	21.16	0.9163	0.4798	44.54	5.28	1094
FLM[16]	18.80	0.8092	0.7672	42.67	5.73	974
LIME[173]	25.02	0.7847	0.5692	40.54	5.43	946
RetinexNet[175]	26.47	0.8500	0.8261	45.10	4.83	1199
Enlighten-GAN[174]	29.21	0.9732	0.8425	**50.20**	5.06	**709**
MIRNet[144]	28.21	0.9782	0.8375	43.55	5.15	1033
FCN[166]	28.63	0.9819	0.8269	43.79	4.78	853
Uformer[154]	28.29	0.9786	0.8324	43.46	5.03	887
TCFA	**30.46**	**0.9870**	**0.8501**	46.37	**4.54**	747

superiority. Compared to the current state-of-the-art approach Uformer, our method improves the quality performance, especially in uneven illumination situations indicating

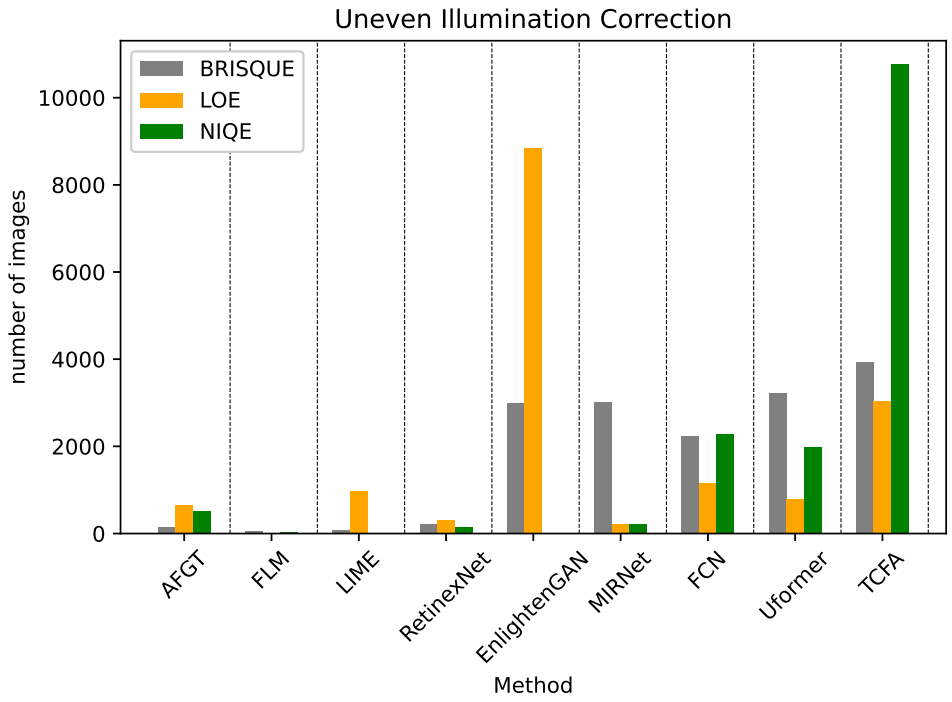
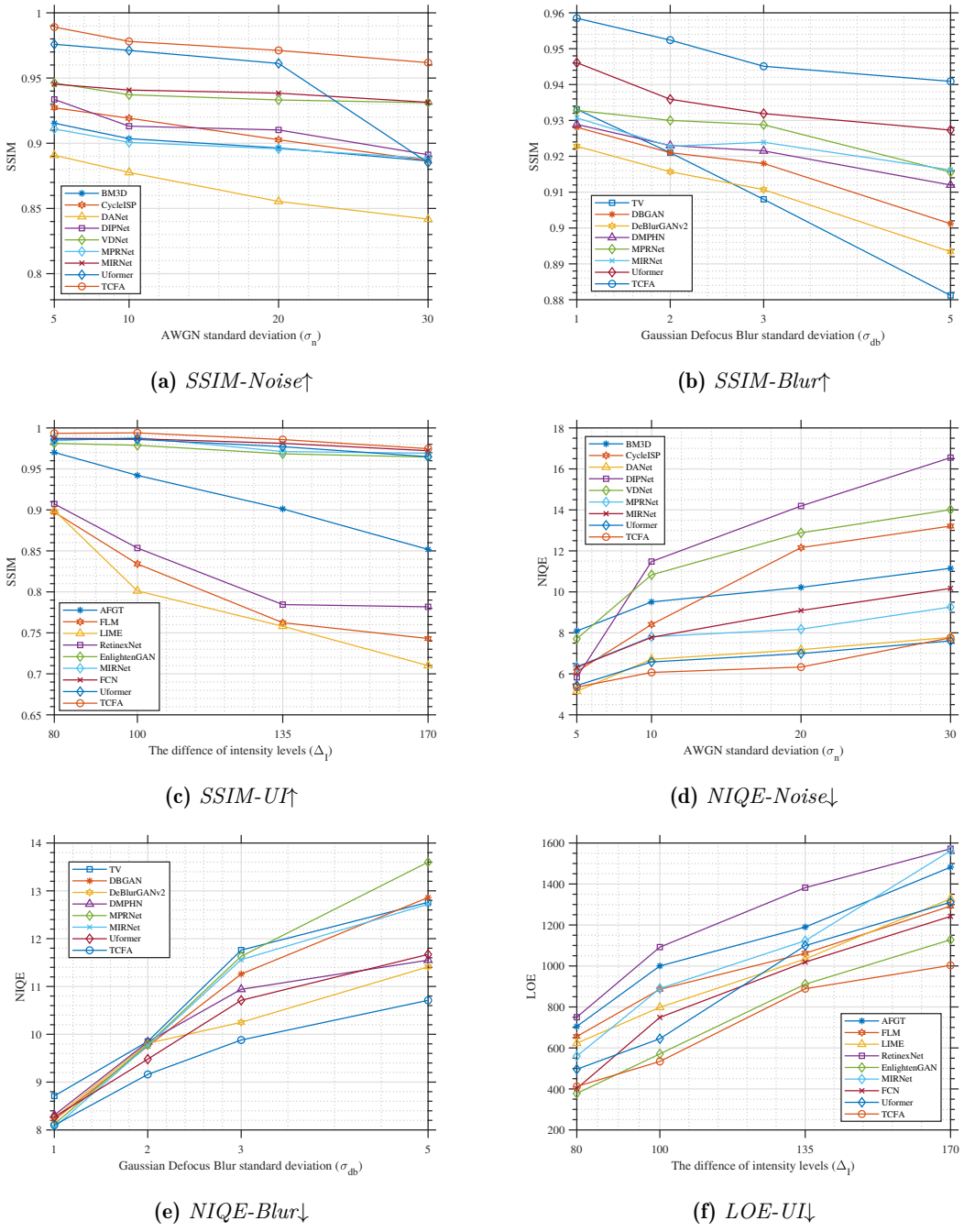


Figure 5.6: *Quantitative comparison with state-of-the-art uneven illumination correction methods in terms of the quantity test on BRISQUE/NIQE/LOE.*

that our model is able to effectively exploit the relationship between distorted image feature and the distortion level features which boost the attention performance.

Since neural networks are well known to be very efficient when they are tested in different levels of distortion (blind enhancement), we propose to evaluate the performance of different enhancement methods with respect to the degraded level. For this reason, for each kind of distortion, images affected by each of the four levels are separately tested and compared.

Fig. 5.7 illustrates the performance of various methods on different quality assessment metrics at varying levels of distortion. As previously mentioned, these plots serve to confirm the high efficiency and effectiveness of recent deep-learning methods such as Uformer[154], EnlightenGAN[174], and MIRNet[144] in comparison to earlier methods like LIME[173] and CycleISP[142] in extracting deep features from original images. Moreover, these plots demonstrate the impact of increasing distortion levels on the performance of these methods. It is evident that the aforementioned CNN-based methods exhibit a significant drop in performance with increasing distortion levels. However, recent methods such as Uformer[154], EnlightenGAN[174], MIRNet[144], as well as the proposed method appear to be less sensitive to distortion levels. Importantly, the results



(a) *SSIM-Noise*↑ **(b)** *SSIM-Blur*↑

(c) *SSIM-UI*↑ **(d)** *NIQE-Noise*↓

(e) *NIQE-Blur*↓ **(f)** *LOE-UI*↓

Figure 5.7: *Effect of the distortion level on the performances for different enhancement methods. The first column shows a performance comparison among different denoising methods. The second column highlights the performance of deblurring methods, while the third column presents the comparative results for correcting uneven illumination.*

also indicate that our proposed TCFA method outperforms state-of-the-art approaches and exhibits greater robustness to serious degradation.

4.3.2 Visual Quality Comparison

As shown in Fig. 5.8,5.9,5.10, which presents a visual illustration of the reconstructed results with various methods, TCFA's results are closest to the reference images. Indeed, our method outperforms others in some aspects. Firstly, as we can observe in the first row of denoising, our approach is better able to recover image details, which can be attributed to the high PSNR achieved by minimizing the reconstruction loss. In contrast, other methods tend to over-enhance images, leading to unexpected blurring. Secondly, as shown in the second row of deblurring, our network produces visually clearer results compared to other methods, which are only able to partially remove the blurry effect.

Furthermore, when it comes to handling uneven illumination, it has been observed that certain low-light enhancement methods such as LIME, and RetinexNet tend to produce over-exposure artifacts, which may even exhibit missing information. Meanwhile, the results obtained by Uformer are generally darker compared to other methods. The results from EnlightenGAN, FCN and MIRNet are also unsatisfactory in producing visually pleasing results in terms of both brightness and naturalness. On the other hand, TCFA is able to successfully enhance the dark areas while preserving the texture details and avoiding over-exposure artifacts. This visualization underscores the drawbacks associated with many low-light enhancement methods and highlights the potential of TCFA as a superior solution.

4.3.3 Statistical Comparison

The box chart compares the distribution of quality assessment scores for different enhancement methods on denoising, deblurring and uneven illumination. Each box represents the Interquartile Range (IQR) of each metric for each method, with the median score indicated by a line within the box. The whiskers extend to the highest and lowest values within 1.5 times the IQR, while outliers beyond this range are plotted as individual points. In Figure 5.11, the detailed statistics of the number of outlier points are summarized to provide insight into the data. Additionally, for a comprehensive comparison, the statistics of reference images (Ref.I.) are presented based on No-Reference IQA metrics.

As presented in the box plots shown in Fig. 5.11a and Fig. 5.11d, it is evident that there is a significant variation in the denoising performance among the compared methods. Especially, the median SSIM score of TCFA was substantially higher (0.98) compared to the other methods, whose median scores ranged from 0.89 to 0.93. Additionally, the IQR of DANet[143] was considerably lower than that of the other methods, indicating

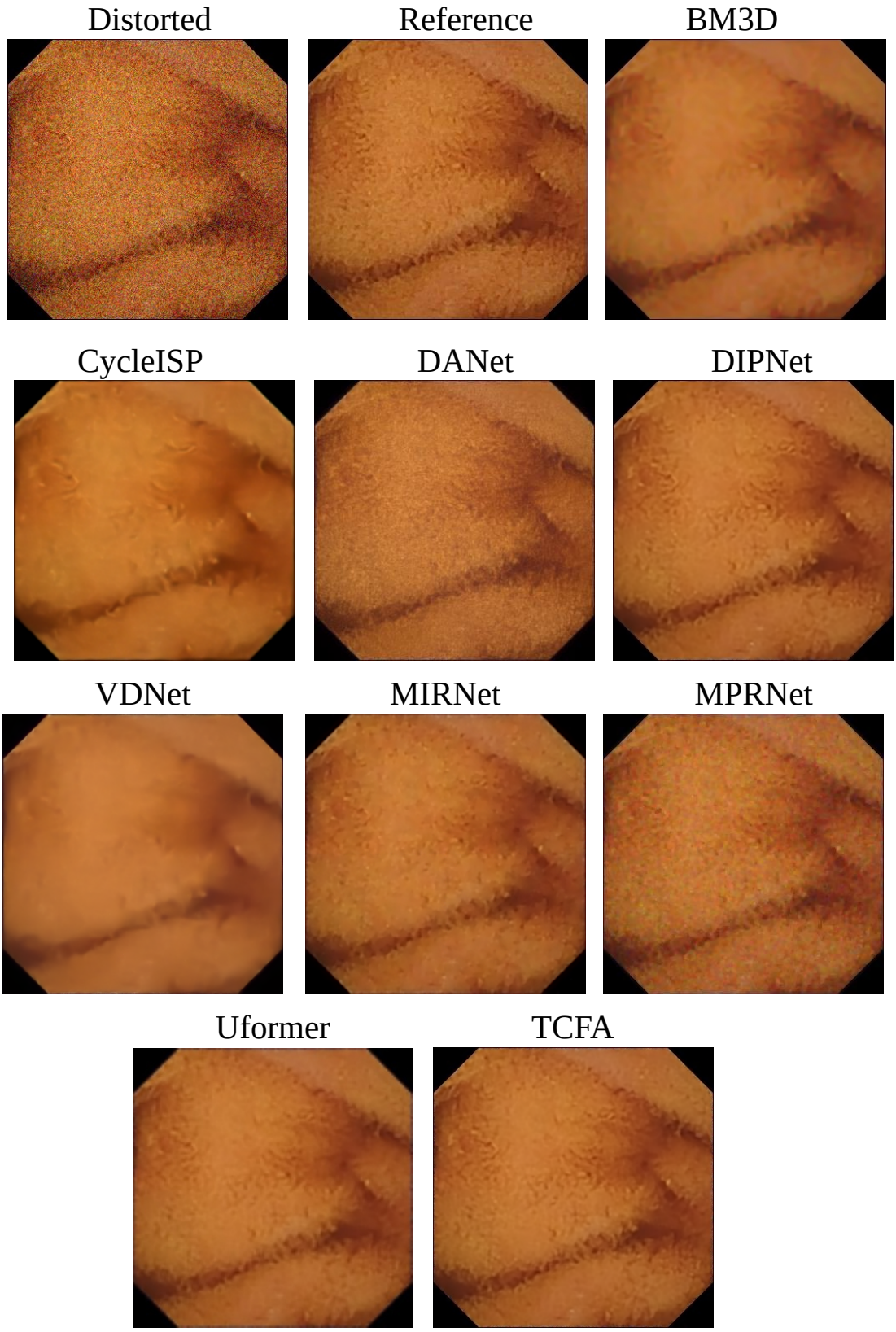


Figure 5.8: *Visual comparison of the reconstruction methods of denoising.*

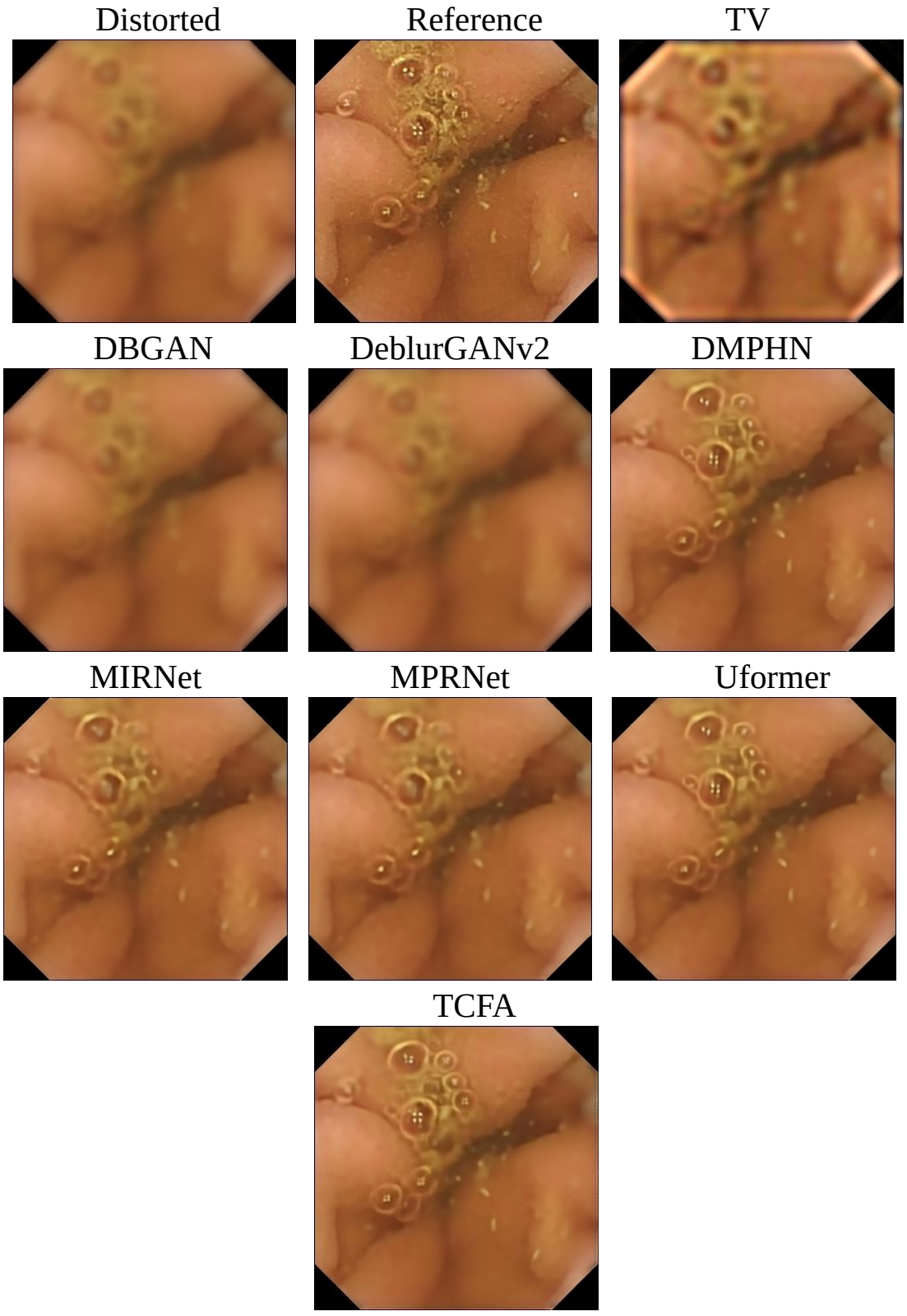


Figure 5.9: *Visual comparison of the reconstruction methods of deblurring.*

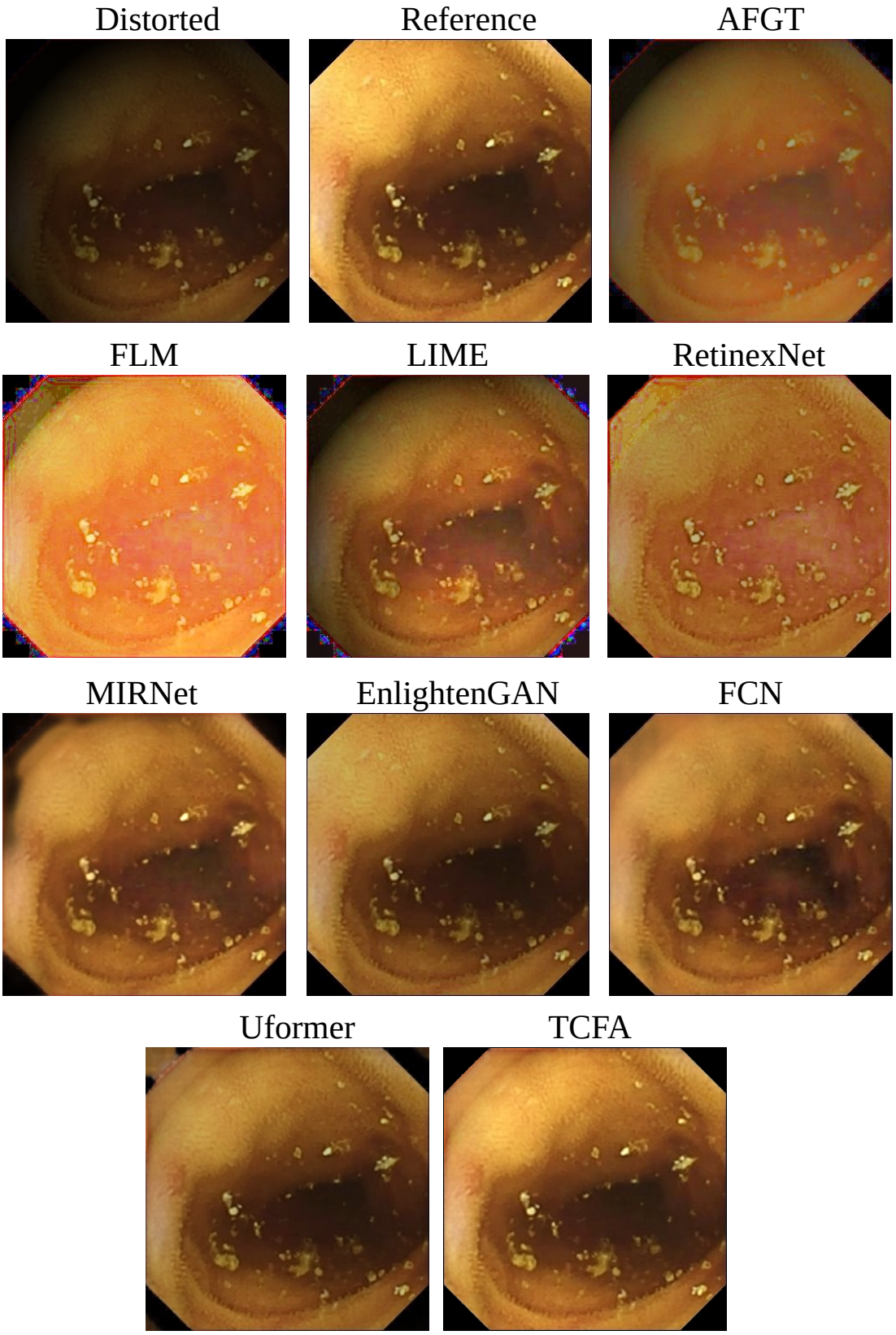
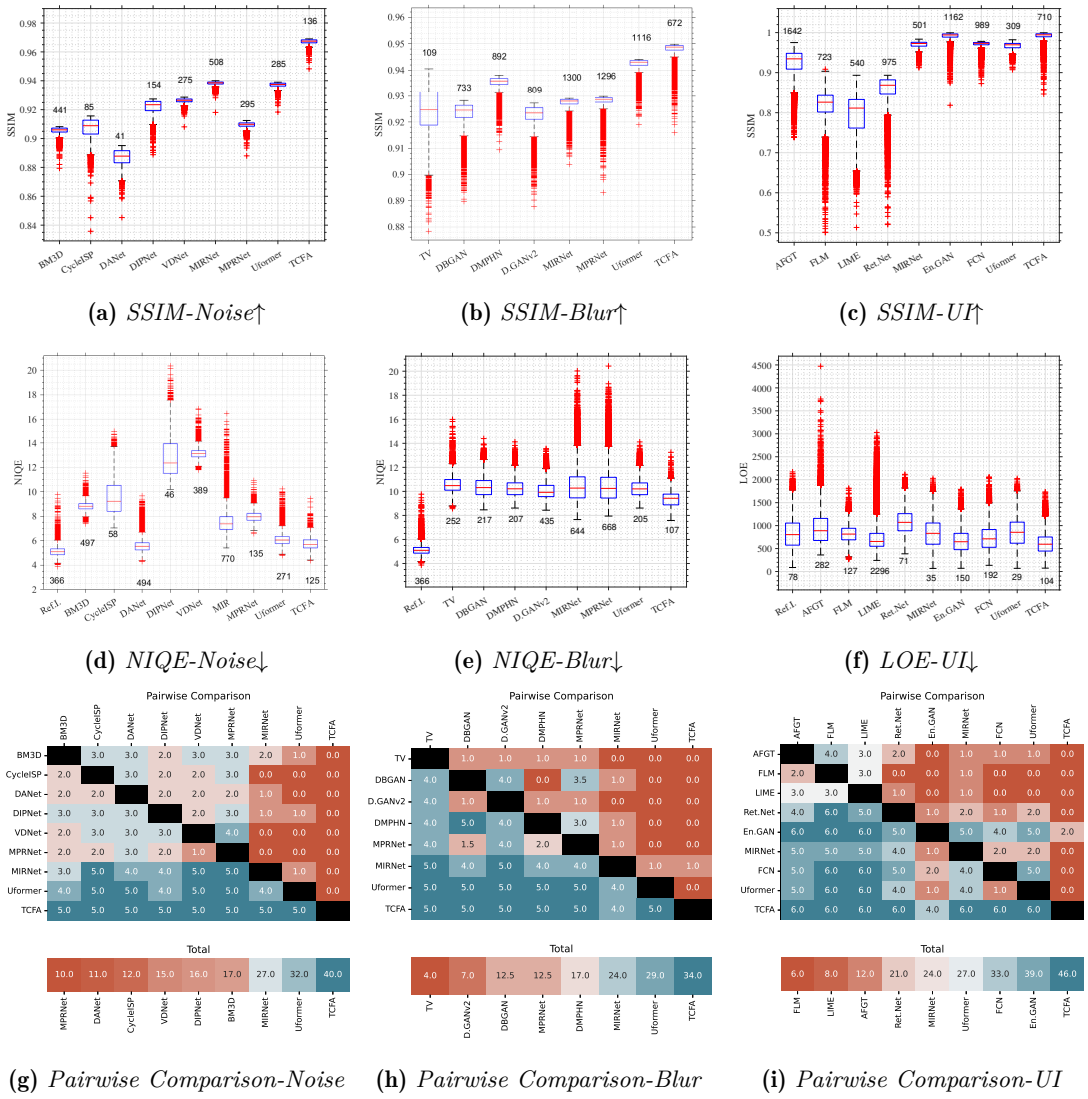


Figure 5.10: *Visual comparison of the reconstruction methods of uneven illumination correction.*



(a) *SSIM-Noise*↑ **(b)** *SSIM-Blur*↑ **(c)** *SSIM-UI*↑

(d) *NIQE-Noise*↓ **(e)** *NIQE-Blur*↓ **(f)** *LOE-UI*↓

(g) *Pairwise Comparison-Noise* **(h)** *Pairwise Comparison-Blur* **(i)** *Pairwise Comparison-UI*

Figure 5.11: *Statistical Comparison of different enhancement methods. The first column shows a comparison among different denoising methods. The second column highlights the performance of deblurring methods, while the third column represents the comparative results for correcting uneven illumination. The first row and the second row show the box plots of the results from different enhancement methods while the third row illustrates the pairwise comparison of the corresponding method on different IQA metrics. In each box plot, the number of outliers related to each considered method is also presented.*

that it was less efficient in its performance. On the other hand, MIRNet[144] exhibited the widest IQR and the most outliers, implying that its performance was less predictable. Overall, these results clearly highlight the substantial difference in performance among the various denoising methods, with TCFA demonstrating the most effective denoising performance for WCE images.

The box plots generated in Fig. 5.11b and Fig. 5.11e revealed that TCFA blurring

had the highest median score compared to the other tested methods. In addition, TCFA blurring had a narrow IQR and few outliers, indicating that it performed consistently across a majority of the quality metrics. On the other hand, Uformer[154] had a similar median score to TCFA blurring but exhibited a wider IQR and more outliers, implying that its performance may not be as reliable across diverse image types and scenarios. Notably, both TCFA blurring and Uformer outperformed the other tested methods, suggesting that they may be the optimal choices for blurring images.

In the context of uneven illumination correction, low-light enhancement techniques were also evaluated to assess their effectiveness in addressing the problem. Results from Fig. 5.11c and Fig. 5.11f revealed that deep-learning-based low-light enhancement methods outperformed two traditional techniques for uneven illumination correction. This finding highlights the potential of deep-learning techniques in addressing issues related to uneven illumination and suggests that they may be a promising alternative for uneven illumination correction. Additionally, it is worth noting that while EnlightenGAN demonstrated competitive performance in terms of median values, the performance of TCFA consistently ranked among the top three methods with the best performance. This finding suggests that while EnlightenGAN may offer a viable option for certain applications, TCFA may be the more reliable choice overall.

Nonetheless, when evaluating multiple methods using individual image quality assessment (IQA) metrics, it is unsatisfactory to conclusively establish the overall superiority of any one method. This is because each method can be optimized based on the specific metric being evaluated. Therefore, further investigation and evaluation are required to fully assess and compare the performance of these methods in various scenarios and contexts. Due to the drawbacks of evaluating methods using singular IQA metrics, we have selected the pairwise comparison [286] as a more comprehensive alternative for this study. This statistical test allows for a head-to-head comparison of two methods by analyzing their performance across a range of IQA metrics. As a result, we can produce a concise and dependable comparison of all considered methods, covering multiple IQA metrics. This will enable us to obtain a more meticulous comprehension of the advantages and limitations of each method, and consequently, derive more robust conclusions regarding their overall performance.

To conduct pairwise comparisons, it is important to establish comparison labels. While there is no definitive guideline for determining the number of labels or points to be used, increasing the number of differentiation options can lead to greater difficulty for the evaluator in justifying their placement. In this study, we have opted to use a

3-point scale as shown in Table 5.4:

Table 5.4: *Pairwise Comparison technique - conversion scale.*

Ratings	Description	Conversion	Inverse
BT	Row's method exhibits **better** performance compared to the column's concerning one IQA metric.	1	WT
AGA	Row's method exhibits the **same** performance compared to the column's concerning one IQA metric.	1/2	AGA
CWT	Row's method exhibits **worse** performance compared to the column's concerning one IQA metric.	0	BT

Fig. 5.11g, 5.11h, and 5.11i shows the pairwise comparison results with respect to denoising, deblurring, and uneven illumination correction, respectively. It is evident that the TCFA method displays superior performance in comparison to other methods across several IQA metrics. This substantiates the conclusion that TCFA is a promising approach for image quality enhancement, given its consistently high-performance scores across various metrics. This finding carries significant implications, suggesting that TCFA could potentially serve as a robust candidate for practical applications in the field of image processing and analysis.

5 Conclusion

In summary, the novel Triplet Clustering Fusion Autoencoder (TCFA) method presented in this study serves as an effective solution to elevate the quality of WCE images across various distortion levels, encompassing noise, blur, and uneven illumination. By adeptly categorizing the severity levels of each distortion type and adeptly extracting both local and global information from the degraded images, the TCFA method demonstrates a substantial enhancement in the performance of image enhancement tasks. Altogether, the TCFA method represents a promising avenue for ameliorating the quality of WCE images, with the potential to significantly bolster the accuracy and efficacy of medical diagnosis and treatment. With image quality enhanced, the subsequent chapter will delve into the classification phase, further advancing the capabilities of this approach in the domain of medical image analysis.

CHAPTER 6

The proposed image classification method: *Dilated Window-based Self-Attention Self-Supervised Learning for Classification in Wireless Capsule Endoscopy*

Abstract

The recent development of Wireless Capsule Endoscopy (WCE) has seen a surge in interest in computer-aided and vision-based solutions, aiming to automate the detection of abnormalities within imagery. While these developments hold great promise, one of the fundamental hurdles in building an effective computer-aided diagnostic (CAD) system for WCE lies in the limited availability of labeled data. To address these complexities and reduce the labeled-data requirement, we propose a novel self-supervised learning method, Dilated Window-based Self-Attention Self-Supervised Learning (DWSA-SSL), tailored for distinguishing various WCE pathologies. Recognizing the challenges posed by the inter-class similarity and intra-class variance in the WCE dataset, we introduce the Shifted-Dilated Window-based Self-Attention (S-WDSA) module within a Transformer-based architecture. This module generates attention maps that encompass not only image patches within the attention window but also their interactions with surrounding windows. The contributions outlined in this study have a profound impact on the enhancement of image classification tasks, as substantiated by a series of extensive experiments conducted on WCE images sourced from the Kvasir-Capsule dataset. Our proposed method undergoes rigorous evaluation on images characterized by significant class imbalance, showcasing its remarkable performance compared to several state-of-the-art approaches.[1]

Chapter content

1 Introduction 122

2 Related Works 124

3 Basic concept of Self-Supervised Learning 127

4 Proposed Method 131

4.1 Algorithm 132

4.1.1 Training phase 133

4.1.2 Testing phase 134

4.1.3 Loss function 134

4.2 S-DWSA Transformer backbone 136

4.2.1 S-DWSA Transformer block 136

4.2.2 Shifted-Dilated Window-based Self-Attention 138

5 Experimental results 140

5.1 Experimental setting 140

5.1.1 Optimization 140

5.1.2 Dataset 141

5.1.3 Evaluation metrics 143

5.2 Ablation study 144

5.2.1 Ablation on hyper-parameters 145

5.2.2 Ablation on scale-soft-cosine attention 147

5.2.3 Ablation on the Shifted-Dilated Window-based Self-Attention 149

5.3 Comparison with state-of-the-art methods 150

6 Conclusion 152

[1]T. -S. Nguyen, J. Chaussard, M. Luong, A. Beghdadi, H. Zaag and T. Le-Tien *"Dilated Window-based Self-Attention Self-Supervised Learning for Classification in Wireless Capsule Endoscopy,"* IEEE Transactions on Medical Imaging (Under preparation for submission).

1 Introduction

Gastrointestinal (GI) cancers constitute a significant portion of digestive system cancers, and became a prevailing concern for public health policies. In 2020, the United States recorded over 300k fresh instances of digestive system cancer diagnoses, with GI tract cancers representing a substantial portion. Correspondingly, an estimated 80k individuals might have succumbed to GI tract cancers during the same year [287].

Wireless capsule endoscopy (WCE) [288] has emerged as a highly effective tool for investigating disorders within the gastrointestinal tract and enables painless imaging of the intestine. However, the widespread applicability and adaptation of WCE face significant challenges related to factors such as efficiency, tolerance, safety, and performance. Moreover, the manual analysis of the extensive dataset generated by WCE presents a time-consuming task requiring specialized skills from medical professionals [289]. Consequently, there has been a growing interest in developing CAD solutions to address these concerns and enable automated analysis for the detection of abnormalities.

In the area of medical image analysis, one of the key challenges of developing a CAD system is the availability of labeled data, particularly when it comes to WCE images [290]. The process of manually annotating and labeling large datasets is not only time-consuming but also requires expertise and domain knowledge from medical professionals. Consequently, the limited availability of labeled WCE images poses a significant obstacle to training accurate and reliable models. Therefore, the use of unlabelled data holds great promise. By effectively harnessing the untapped information contained within unlabelled WCE images, it becomes possible to train models that are better equipped to detect abnormalities, classify different pathologies, and support medical practitioners in their diagnostic tasks.

Recently, the exploration of semi-supervised learning (SSL) [291] techniques within the domain of WCE image processing presents an opportunity to overcome the limitations imposed by the scarcity of labeled data. While semi-supervised learning can be advantageous in certain scenarios, it also has some drawbacks when applied to WCE image classification. Indeed, WCE datasets with labeled data are often scarce and expensive to obtain, as they require expert annotations. This scarcity of labeled data restricts the effectiveness of semi-supervised learning, as it heavily relies on a small set of labeled examples to guide the learning process. Insufficient labeled data can lead to sub-optimal performance and limited generalization. Besides, semi-supervised learning algorithms usually make assumptions about the underlying data distribution. More precisely, to efficiently generate the pseudo label for unlabeled data, the labeled data

has an assumption to be clustered. In WCE image classification, these assumptions may not hold true due to the complex and diverse nature of gastrointestinal images. The inter-class similarity and intra-class variance in the WCE dataset play a challenge in generating discriminative features for pseudo-label learning. If the labeled data is not representative of the entire data distribution or contains biases, the classifier may learn and propagate those biases, resulting in biased predictions. Moreover, unlabeled data, which is abundant in semi-supervised learning, may contain noise, outliers, or irrelevant samples. When incorporating this noisy data into the training process, the classifier can inadvertently amplify the impact of the noise, leading to erroneous predictions. Noise in the unlabeled data can adversely affect the performance and reliability of the WCE classification model.

To deal with these labeled-data requirement difficulties, we propose a novel Dilated Window Self-Attention Self-Supervised Learning (DWSA-SSL) method to identify by classification from WCE images, various GI pathologies/findings, namely erythema, angiectasias, blood-fresh, erosion, ulcers, lymphangiectasia, polyp, and normal clean mucosa. Besides, the inter-class similarity and intra-class variance in the WCE dataset present a challenge in generating discriminative features for classification. In fact, the regions around the lesion can be considered as a joint criterion to classify the presented pathology. From this observation, we propose the Shifted-Dilated Window-based Self-Attention (S-WDSA) in the Transformer-backbone which generates the attention map between not only the image patches in the attention window but also in-cooperated surrounding windows.

The training phase of the proposed two-branch SSL method is performed with two main tasks including the pretext task and the downstream task. In the pretext task, known as the upstream task for pre-training, both online and target branches are used to guide the model to learn intermediate representations of unlabelled WCE images. The target is only updated from the knowledge transfer process of the online branch. Once the pretext task, the downstream task training is processed only on the target model to the specific WCE classification task. Downstream tasks are provided with less quantity of labeled WCE images. In the testing phase, the target model is evaluated with the WCE image testing set.

The main contributions of our work can be summarized in three points:

- We propose a comprehensive asymmetric SSL network structure for WCE image classification. It is successful in understanding the underlying structural meaning of WCE images which is beneficial for the practical downstream task in the second

branch. The ability to train SSL models with unlabeled WCE images speeds up the overall training process and empowers the model to learn underlying semantic features without introducing label bias.

- We propose to use the transformer-backbone to reduce the redundant and noisy information for image classification thus avoiding biased performance. Transformer can highlight the most important regions in images. Therefore, it is beneficial to introduce the attention mechanism to focus on lesions while making diagnosis predictions. Moreover, the proposed S-WDSA module generates a hybrid attention embedding that is based on both image patches in the attention windows and between in-cooperated surrounding windows. By integrating information from both local and global contexts, the S-WDSA module enriches our model's understanding of the image, thereby enhancing its ability to make accurate and contextually informed diagnostic predictions.

- In self-attention, the learned attention maps of some blocks and heads are frequently dominated by a few pixel pairs which makes the attention map biased to some specific region. To ease this case, we propose a scale-soft-cosine attention approach that computes the attention logit of the pixel pair by a soft-cosine function which results in the attention map value being much milder than in the original configuration.

The subsequent sections of this work are structured as follows. Section 2 presents the recent related classification methods applied on WCE. Following this, Section 3 covers the fundamental concept of self-supervised learning while also engaging in a comprehensive discussion of several established self-supervised methods that have notably influenced and inspired the development of our proposed approach. Subsequently, Section 4 provides a detailed account of the proposed method. After that, Section 5 furnishes the experimental results, encompassing the ablation study and a thorough comparison with state-of-the-art methods. These results offer empirical validation of the efficiency of our proposed approach, further substantiating its contributions to the field. Finally, this chapter concludes in Section 6, summarizing the key findings and implications of this work.

2 Related Works

Significant efforts have been directed towards the development of robust and efficient methods for classifying abnormalities in Wireless Capsule Endoscopy (WCE) images.

These endeavors can be broadly categorized into three primary groups: traditional machine learning techniques, deep learning-based methods, and more recent attention-based approaches.

Within the domain of conventional machine learning-based methods, diverse techniques have been explored and employed for a range of tasks, including classification, regression, and clustering. In a study by Sekuboyina et al. [292], the initial step involves the conversion of endoscopy images into CIE-Lab and YCbCr color spaces. Subsequently, the a-channel of the CIE-Lab space serves as input for a convolutional neural network (CNN) to conduct classification. To tackle the challenge of imbalanced data, the Synthetic Minority Over-sampling Technique (SMOTE) is implemented as a method for oversampling to address the imbalance issue. Sindhu et al. [293] initially extracted conventional SIFT features in conjunction with Haralick texture features. These two sets of features are then merged to define the characteristics of image patches, which are subsequently fed into a Neural Network (NN) for supervised classification.

In a related study carried out by Sharif et al. [294], a combination of geometric attributes and CNN-derived features is employed. The geometric attributes are generated through the transformation of the image into the CIELab color space, followed by the application of a thresholding technique. Concurrently, CNN features are extracted using the VGG16 and VGG19 networks. These combined features are then used as input for K-nearest neighbor classification.

Working toward the achievement of Deep Learning-based methods' efficacy, which has been substantiated in multi-pathology classification [225], [295], the procedure involves extracting texture attributes. These attributes encompass features such as local binary patterns and haralick features, in addition to more traditional descriptors like joint composite descriptor, auto color correlogram, color layout, and edge histogram. These various features are amalgamated and input into a CNN for the classification task.

Lan et al. [226] introduce a novel deep cascade network known as cascade-proposal, tailored for multi-pathology classification. Their method encompasses two primary components: firstly, the generation of a limited set of region proposals with high recall, achieved via a region proposal rejection module; in parallel, the identification of abnormal patterns through a detection module. To pinpoint regions of interest, they employ the multi-regional combination (MRC) technique, and for precision in region localization, they use the salient region segmentation (SRS) method. Additionally, they integrate the dense region fusion (DRF) method to enhance the refinement of object boundaries.

Sadasivan et al. [222] tackle the challenge of multi-pathology classification by segmenting WCE images into multiple patches. Each patch is characterized based on extracted color and texture attributes derived from the chromatic components found within the CIE Lab color space. To categorize these patches, K-means clustering is applied to group the extracted attributes into distinct clusters, followed by the utilization of dictionary learning for classification purposes.

In recent years, attention techniques have emerged as a prominent approach for improving multi-pathology classification tasks. Attention mechanisms enable models to selectively focus on important regions or features within an image, enhancing overall classification performance by highlighting relevant information while suppressing irrelevant or noisy signals. Several recent studies have explored the application of attention mechanisms in this context. Xing et al. [227], [296] proposed a method that incorporates attention mechanisms and deformation to zoom in on lesions. They extract attention maps using self-attention and then use deformation techniques to zoom in on the identified lesion region. By comparing the self-attention maps before and after deformation, they generate a final attention map for classification.

To address the challenge of value bias in attention loss, Guo and Yuan [297], [298] introduced the use of angular loss for classification. They applied attention mechanisms to generate different attention maps of multi-level features. Additionally, they incorporated a semi-supervised approach to handle unlabeled data. In a similar vein, Zhao et al. [299] employed multiple attention layers and used the adaptive cosine loss for classification refinement. By incorporating attention mechanisms, they aimed to enhance the discriminative power of the model.

These studies highlight the effectiveness of attention techniques in multi-pathology classification. Attention mechanisms allow for selective focus on informative regions or features, thereby improving classification accuracy and robustness. Additionally, to address specific challenges such as value bias and unlabeled data, innovative approaches such as self-supervised learning could be considered. For this reason, we propose a new approach dilated window self-attention SSL which learns from the data itself. We can train the proposed model not only to look at the main problem area in the image but also to consider the nearby parts. By doing so, the model not only understands the main features but also learns how it's related to the surrounding parts.

Using self-supervised learning is like having a two-in-one benefit. It helps the model pay attention to the right things and also learn from the context around those things. This double learning helps us improve how we classify different diseases. It's like having

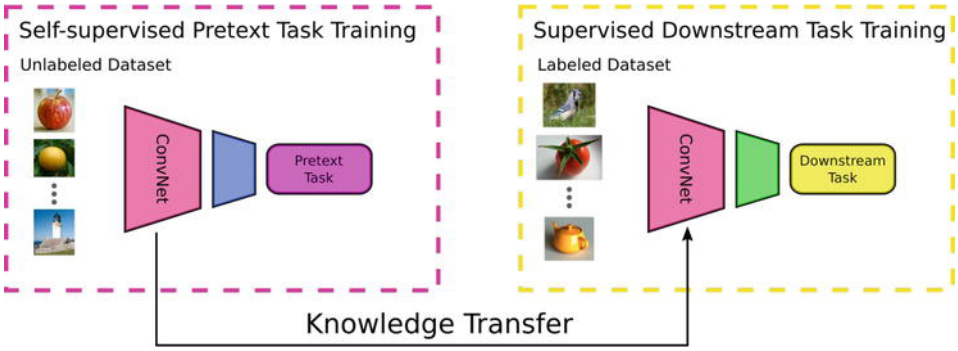


Figure 6.1: *A model is first trained with a pretext task with unlabeled data, then fine-tuned on the downstream task with a limited amount of labeled data. Usually, convolution layers, which are mostly responsible for learning representations, are transferred. A few fully connected layers towards the end are changed or retrained.*

a better map of what's happening in the picture, making our predictions more accurate and dependable. Before we explain our method in detail, we'll first talk about the basic idea of SSL in the next section.

3 Basic concept of Self-Supervised Learning

Self-supervised learning allows us to exploit a variety of labels that come within the data itself. Producing a dataset with clean labels is expensive but unlabeled data is being generated all the time. To make use of this much larger amount of unlabeled data, one way is to set the learning objectives properly so as to get supervision from the data itself.

Since a self-supervised model does not know the actual labels corresponding to the inputs, its success depends on the design of the pretext tasks to generate the pseudo-labels from part of the input data. Final performance on the pretext task is not important, but we hope that the learned intermediate representations can capture good information and be beneficial to a variety of downstream tasks.

Fig. 6.1 illustrates the process of transferring knowledge from self-supervised training to supervised training. Once the training of the pretext task network is done, we save the convolutional layers, which are responsible for generating learned representations. Then, we train a supervised downstream network with a limited amount of labeled data, by adding fully connected layers and placing a classifier head. This concept of knowledge transfer has been embraced by numerous SSL techniques for classification, which have proven their efficacy in acquiring meaningful representations from unlabeled data.

SimCLR: [300] (Chen et al, 2020) introduced an uncomplicated framework for the contrastive learning of visual representations. This method focuses on acquiring representations for visual inputs by enhancing agreement between variously augmented views of the same data sample through the use of a contrastive loss function within the latent space.

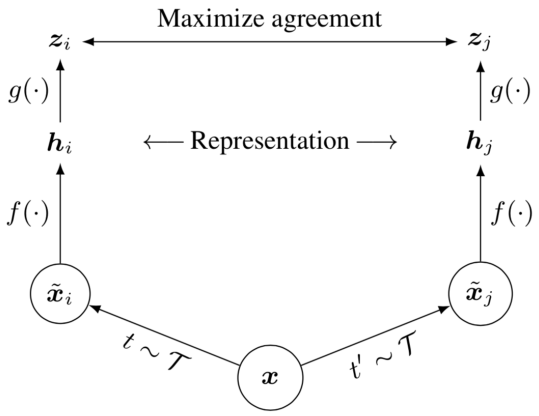


Figure 6.2: *A simple framework for contrastive learning of visual representations.*

i. Randomly sample a minibatch of N samples and each sample is applied with two different data augmentation operations, resulting in $2N$ augmented samples in total.

$$\tilde{\mathbf{x}}_i = t(\mathbf{x}), \quad \tilde{\mathbf{x}}_j = t'(\mathbf{x}), \quad t, t' \sim \mathcal{T} \tag{6.1}$$

ii. Given one positive pair, other $2(N-1)$ data points are treated as negative samples. The representation is produced by a base encoder $f(.)$:

$$\mathbf{h}_i = f(\tilde{\mathbf{x}}_i), \quad \mathbf{h}_j = f(\tilde{\mathbf{x}}_j) \tag{6.2}$$

iii. The contrastive learning loss is defined using cosine similarity $\text{sim}(.,.)$. Note that the loss operates on an extra projection layer of the representation $g(.)$ rather than on the representation space directly. But only the representation $\mathbf{h}$ is used for downstream tasks.

$$\begin{gathered} \mathbf{z}_i = g(\mathbf{h}_i), \quad \mathbf{z}_j = g(\mathbf{h}_j) \\ \mathcal{L}_{\text{SimCLR}}^{(i,j)} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \end{gathered} \tag{6.3}$$

where $\mathbb{1}_{[k \neq i]}$ is an indicator function: $\mathbb{1}$ if $k \neq i$, 0 otherwise.

SimCLR needs a large batch size to incorporate enough negative samples to achieve good performance. Fig. 6.3 presents the learning process of the SimCLR algorithm.

Algorithm 1 SimCLR's main learning algorithm.

input: batch size N, constant τ, structure of f, g, $\mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^{\top} \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i, j)$ **as** $\ell(i, j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

Figure 6.3: *The algorithm for SimCLR [300].*

BYOL: In contrast to the previously mentioned approaches, BYOL (Bootstrap Your Own Latent) [301] presents an intriguing deviation. Remarkably, BYOL asserts that it attains cutting-edge performance without the need for negative samples. Instead, it utilizes two neural networks, namely the online and target networks, which collaborate and mutually learn. The target network, characterized by parameter ξ, shares the same architecture as the online network, parameterized by θ, but with weights smoothed using the Polyak averaging technique, represented as $\xi \leftarrow \tau\xi + (1-\tau)\theta$.

The primary objective is to acquire a representation denoted as "y" that can be applied in subsequent tasks. The online network, defined by the parameter θ, comprises three key components: an encoder denoted as f_θ, a projector called g_θ, and a predictor referred to as q_θ.

The target network shares an identical architectural structure with the online network but employs a distinct parameter, denoted as ξ. The parameter ξ is updated using Polyak averaging with the θ parameter, as expressed by the equation: $\xi \leftarrow \tau\xi + (1-\tau)\theta$. Given an image $\mathbf{x}$, the BYOL loss is constructed as follows:

i. Create two augmented views: $\mathbf{v} = t(\mathbf{x}); \mathbf{v}' = t'(\mathbf{x})$ with augmentations sampled

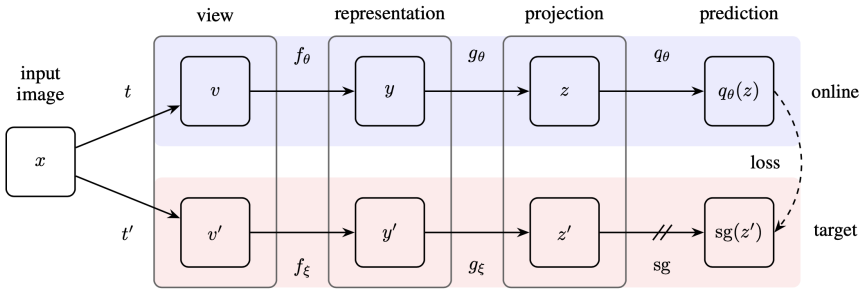


Figure 6.4: *The model architecture of BYOL. After training, we only care about f_θ producing representation, $y = f_\theta(x)$, and everything else is discarded. sg means stop gradient.*

$$t \sim \mathcal{T}, t' \sim \mathcal{T}'$$

ii. Then they are encoded into representations: $\mathbf{y}_\theta = f_\theta(\mathbf{v}), \mathbf{y}' = f_\xi(\mathbf{v}')$

iii. Then they are projected into latent variables: $\mathbf{z}_\theta = g_\theta(\mathbf{y}_\theta), \mathbf{z}' = g_\xi(\mathbf{y}')$

iv. The online network outputs a prediction: $q_\theta(\mathbf{z}_\theta)$

v. Both $q_\theta(\mathbf{z}_\theta)$ and $\mathbf{z}'$ are L2-normalized, giving us: $\bar{q}_\theta(\mathbf{z}_\theta) = q_\theta(\mathbf{z}_\theta)/|q_\theta(\mathbf{z}_\theta)|$

vi. The loss $\mathcal{L}_\theta^{\mathrm{BYOL}}$ is MSE between L2-normalized prediction $\bar{q}_\theta(\mathbf{z})$ and $\bar{\mathbf{z}}'$

vii. The other symmetric loss $\tilde{\mathcal{L}}_\theta^{\mathrm{BYOL}}$ can be generated by switching $\mathbf{v}'$ and $\mathbf{v}$; that is, feeding $\mathbf{v}'$ to online network and $\mathbf{v}$ to target network.

viii. The final loss is $\mathcal{L}_\theta^{\mathrm{BYOL}} + \tilde{\mathcal{L}}_\theta^{\mathrm{BYOL}}$ and only parameters θ are optimized.

In contrast to the majority of widely adopted contrastive learning methods, BYOL takes a unique approach by not utilizing negative pairs. While many bootstrapping techniques depend on pseudo-labels or cluster indices, BYOL directly leverages the latent representation for bootstrapping.

MoCo & MoCo-V2: Momentum Contrast (MoCo [302]) provides a framework of unsupervised learning visual representation as a dynamic dictionary look-up. The dictionary is structured as a large FIFO queue of encoded representations of data samples.

Given a query sample $\mathbf{x}_q$, we acquire a representation for the query by utilizing an encoder, denoted as $\mathbf{q} = f_q(\mathbf{x}_q)$, and then obtain a collection of key representations, $\mathbf{k}_1, \mathbf{k}_2, \ldots$, stored in a dictionary. These key representations are generated through encoding with a separate momentum encoder, $\mathbf{k}_i = f_k(\mathbf{x}_i^k)$. It's worth noting that within this collection, there exists a single positive key, $\mathbf{k}^+$, which corresponds to the query

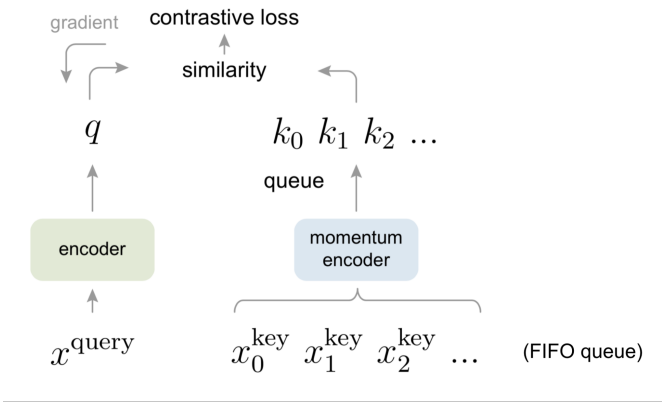


Figure 6.5: *Illustration of how Momentum Contrast (MoCo) learns visual representations.*

$\mathbf{q}$. The authors construct $\mathbf{k}^+$ by creating a duplicate of $\mathbf{x}_q$ with additional variations. Subsequently, they apply the InfoNCE contrastive loss [303], using a temperature parameter denoted as τ, to compare one positive sample against a set of $N-1$ negative samples:

$$\mathcal{L}_{\text{MoCo}} = -\log \frac{\exp(\mathbf{q} \cdot \mathbf{k}^+/\tau)}{\sum_{i=1}^{N} \exp(\mathbf{q} \cdot \mathbf{k}_i/\tau)} \tag{6.4}$$

In MoCo, a dictionary based on a queue system offers us the advantage of reusing representations from the immediately preceding mini-batches of data. It's important to note that the MoCo dictionary does not possess differentiability in the form of a queue, which means we cannot rely on back-propagation to update the key encoder, denoted as f_k. One simplistic approach could involve using the same encoder for both f_q and f_k. However, MoCo takes a distinct approach by suggesting the utilization of a momentum-based update mechanism, incorporating a momentum coefficient denoted as $m \in [0,1)$. To elaborate, let's designate the parameters of f_q and f_k as θ_q and θ_k, respectively.

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \tag{6.5}$$

MoCo's strength, when contrasted with SimCLR, lies in its ability to disentangle batch size from the number of negatives. SimCLR, on the other hand, necessitates large batch size to secure a sufficient pool of negative samples, and it faces a decline in performance when batch sizes are reduced.

4 Proposed Method

In this section, we provide a comprehensive and detailed description of the proposed DWSA-SSL method. This description will focus on the intricacies of the method,

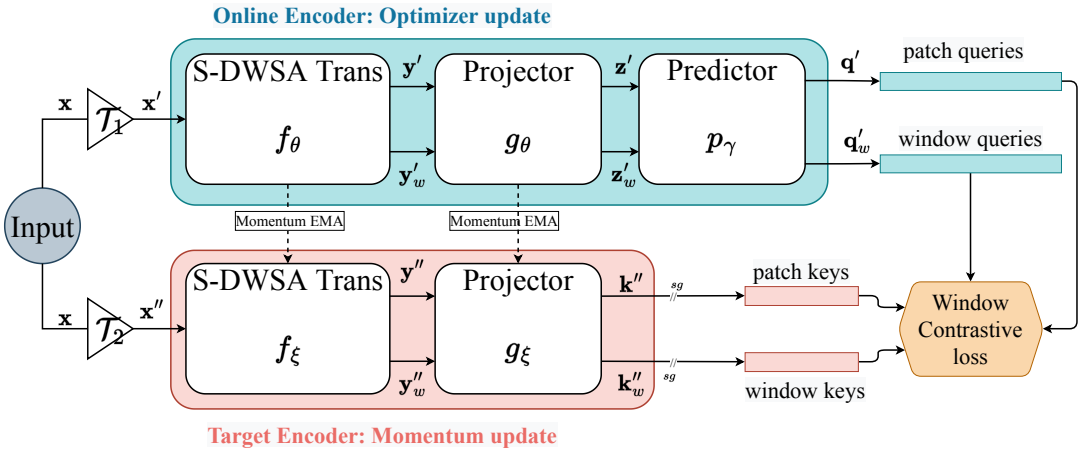


Figure 6.6: *The pipeline of the proposed DWSA-SSL method.* $\mathcal{T}_1, \mathcal{T}_2$ *are image augmentations.* $\mathbf{x}', \mathbf{x}''$ *are two augmented views from* $\mathbf{x}$ *by applying respectively* $\mathcal{T}_1, \mathcal{T}_2$. $\mathbf{y}', \mathbf{y}'_w$ *and* $\mathbf{y}'', \mathbf{y}''_w$ *are the online/ target representations corresponding to the patch-level and window-level attention by S-DWSA Transformer backbones, respectively.* $\mathbf{z}', \mathbf{z}'_w$ *are the online projections of* $\mathbf{y}', \mathbf{y}'_w$, *respectively.* $\mathbf{q}', \mathbf{q}'_w$ *are the online prediction of* $\mathbf{z}', \mathbf{z}'_w$, *respectively.* $\mathbf{k}'', \mathbf{k}''_w$ *are the target projections (key) of* $\mathbf{y}'', \mathbf{y}''_w$, *respectively. EMA: Exponential Moving Average. sg means stop-gradient.*

explaining its components, functionalities, and underlying principles.

Taking inspiration from MoBY [304], our proposed method represents a fusion of two widely recognized self-supervised learning strategies: MoCo v2 [305] and BYOL [301]. It inherits key elements such as the momentum-based architecture, key queue, and the application of the contrastive loss, which are drawn from MoCo v2. Additionally, it adopts characteristics like asymmetric encoders, dissimilar data augmentations, and the momentum scheduling mechanism present in BYOL.

Fig. 6.6 presents an illustration of the proposed method. It comprises two encoders: an online encoder and a target encoder. Each of these encoders is composed of a backbone and a projector head (a 2-layers MLP). Notably, the online encoder introduces an additional prediction head (also a 2-layers MLP), creating an asymmetry between the two encoders. During training, the online encoder receives updates through gradient computations, while the target encoder undergoes adjustments through a moving average process, employing momentum updates at each training iteration.

4.1 Algorithm

Firstly, our proposed approach operates on pairs of distorted images' embeddings. To elaborate, when given a batch of N images represented as $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_N$, we generate

two sets of distorted views denoted as $\mathbf{X}' = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_N\}$ and $\mathbf{X}'' = \{\mathbf{x}'', \ldots, \mathbf{x}''\}$ using two distinct stochastic data augmentation techniques, $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively.

Subsequently, we employ two transformers, denoted as f_θ and f_ξ, along with two projectors, g_θ and g_ξ, each characterized by parameters θ and ξ. These components are used to generate corresponding embeddings, namely $\mathbf{z}'$, $\mathbf{z}'_w$, $\mathbf{k}''$, and $\mathbf{k}''_w$. It's important to note that $\mathbf{z}'$ and $\mathbf{k}''$ represent embeddings associated with attention at the image patch level, whereas $\mathbf{z}''_w$ and $\mathbf{k}''_w$ represent attention at the window level, generated using the Shifted-Dilated Window-based Self-Attention Transformer. Further details regarding the backbone will be provided in Section 4.2.

$$[\mathbf{z}', \mathbf{z}'_w] = g_\theta(f_\theta(\mathbf{x}')) \tag{6.6}$$

$$[\mathbf{k}'', \mathbf{k}''_w] = g_\xi(f_\xi(\mathbf{x}'')) \tag{6.7}$$

To simplify the notation, we assume that the outcomes produced by the projector have been normalized to unit vectors. In order to introduce an asymmetry into the architecture, the predictor, denoted as p_γ, is responsible for generating the queries, namely $\mathbf{q}'$ and $\mathbf{q}'_w$, derived from the embeddings $\mathbf{z}'$ and $\mathbf{z}'_w$.

$$[\mathbf{q}', \mathbf{q}'_w] = p_\gamma([\mathbf{z}', \mathbf{z}'_w]) \tag{6.8}$$

4.1.1 Training phase

The training phase of the proposed two-branch SSL method consists of two primary tasks, the pretext task and the downstream task

i. During the pretext task, which serves as the upstream task for pre-training, both the online and target branches are utilized to guide the model in learning intermediate representations from unlabeled WCE images. The knowledge transfer process from the online branch updates the target branch exclusively.

ii. After completing the pretext task, the downstream task training is conducted solely on the target model, focusing on the specific WCE classification task. The downstream tasks are provided with a smaller quantity of labeled WCE images.

In our setting, we apply the same momentum encoder technique as proposed in MoCo [302], such that only the parameter θ is updated through backpropagation, and the parameters ξ is the exponential moving average of the parameter θ. At time step t, we have:

$$\xi_t = EMA(\theta_t) = \alpha\xi_{t-1} + (1-\alpha)\theta_t \tag{6.9}$$

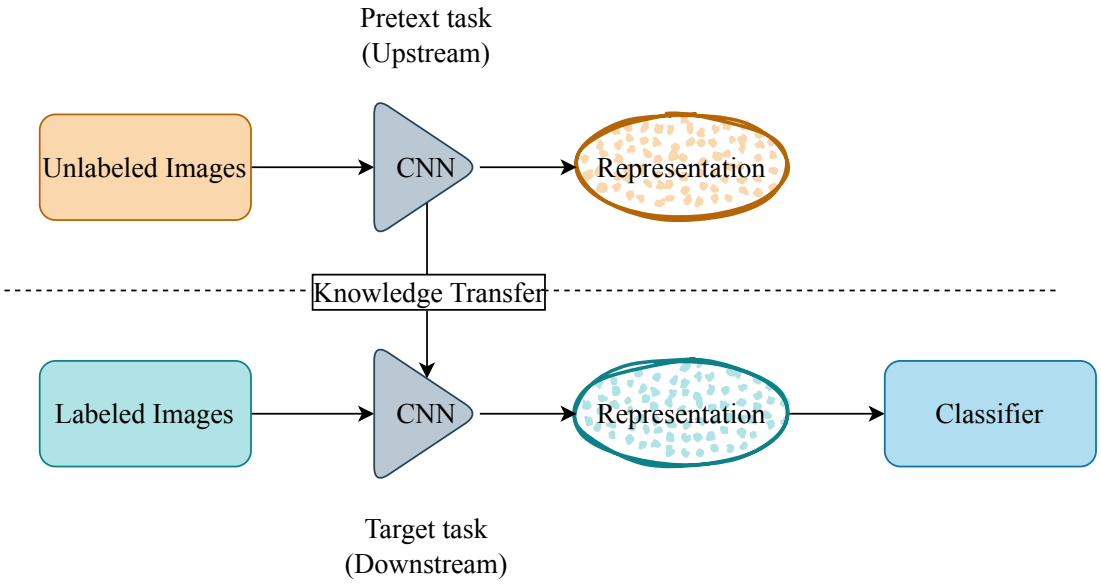


Figure 6.7: *Overview of SSL approaches.*

4.1.2 Testing phase

In the testing phase, the downstream model of self-supervised learning is used to validate the image. This involves employing the trained downstream model, which has been trained using self-supervised learning techniques, to classify the pathologies. The downstream model acts as a validation mechanism, leveraging the knowledge and representations learned during the self-supervised learning phase to make informed judgments about the image's attributes.

4.1.3 Loss function

Intuitively, the loss function simultaneously addresses two objectives. The first component of the loss aims to bring the embeddings of various augmentations of the same image (referred to as the positive pair) closer together. Conversely, the second component endeavors to push the negative pair in opposite directions from each other, ultimately minimizing their exponent. The formulation of this loss function is inspired by the InfoNCE loss [303], which employs a categorical cross-entropy loss to distinguish the positive sample from a set of unrelated noise samples.

More precisely, most recent studies typically adhere to the following definition of a contrastive learning objective, which allows for the incorporation of multiple positive and negative samples. In this context, we have the data distribution denoted as $p_{\texttt{data}}(.)$ over $\mathbb{R}^n$ and the distribution of positive pairs, represented as $p_{\texttt{pos}}(.,.)$, over $\mathbb{R}^{n\times n}$. It's essential that these two distributions adhere to the following conditions:

- Symmetry: $\forall \mathbf{x}, \mathbf{x}^+, p_{\texttt{pos}}(\mathbf{x}, \mathbf{x}^+) = p_{\texttt{pos}}(\mathbf{x}^+, \mathbf{x})$
- Matching marginal: $\forall \mathbf{x}, \int p_{\texttt{pos}}(\mathbf{x}, \mathbf{x}^+) d\mathbf{x}^+ = p_{\texttt{data}}(\mathbf{x})$

To help an encoder $f(\mathbf{x})$ to learn a L2-normalized feature vector, the contrastive learning objective is:

$$
\begin{aligned}
&\mathcal{L}_{\text{contrastive}} = \\
&\mathbb{E}_{\substack{(\mathbf{x},\mathbf{x}^+)\sim p_{\text{pos}}, \\ \{\mathbf{x}_i^-\}_{i=1}^M \overset{\text{i.i.d}}{\sim} p_{\texttt{data}}}} \Big[-\log \frac{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau}}{e^{f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau} + \sum_{i=1}^M e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \Big] \\
&\approx \mathbb{E}_{\substack{(\mathbf{x},\mathbf{x}^+)\sim p_{\text{pos}}, \\ \{\mathbf{x}_i^-\}_{i=1}^M \overset{\text{i.i.d}}{\sim} p_{\texttt{data}}}} \Big[-f(\mathbf{x})^\top f(\mathbf{x}^+)/\tau + \log \big(\sum_{i=1}^M e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau} \big) \Big] \\
&= -\frac{1}{\tau} \mathbb{E}_{(\mathbf{x},\mathbf{x}^+)\sim p_{\text{pos}}} f(\mathbf{x})^\top f(\mathbf{x}^+) \\
&\quad + \mathbb{E}_{\mathbf{x}\sim p_{\texttt{data}}} \Big[\log \mathbb{E}_{\mathbf{x}^-\sim p_{\texttt{data}}} \big[\sum_{i=1}^M e^{(f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau)} \big] \Big]
\end{aligned}
\tag{6.10}
$$

where M is the number of negative samples. Therefore, the InfoNCE can be written into:

$$
\mathcal{L}_{\text{InfoNCE}} = \frac{-1}{N} \sum_{i=1}^N f(\mathbf{x})^\top f(\mathbf{x}^+) + \frac{\tau}{N} \sum_{i=1}^N log \sum_{j=1}^M e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau} \tag{6.11}
$$

Given that the lesion can be represented not only within the attention window but also identified by the neighboring regions, it becomes imperative to enhance the efficiency of its representation. By capturing the spatial relationships and dependencies among adjacent regions, we propose a novel Window-based Contrastive Loss $\mathcal{L}_{\text{WC}}$ as:

$$
\begin{aligned}
\mathcal{L}_{\text{WC}}(\mathbf{q}', \mathbf{q}'_w, \mathbf{k}'', \mathbf{k}''_w) = &\frac{-1}{N} \sum_{i=1}^N \frac{\mathbf{q}_i'^\top \mathbf{k}_i''}{cos(\mathbf{q}'_{w-i}, \mathbf{k}''_{w-i})} \\
&+ \frac{\tau}{N} \sum_{i=1}^N log \sum_{j=1}^M e^{\left(\frac{\mathbf{q}_i'^\top \mathbf{k}_j''}{cos(\mathbf{q}'_{w-i}, \mathbf{k}''_{w-j})} / \tau \right)},
\end{aligned}
\tag{6.12}
$$

where $cos(.)$ denotes the cosine similarity. The occurrence of inter-class similarity and intra-class variance can be attributed to the resemblance observed in the surrounding regions of the lesion. Consequently, using cosine similarity enables the model to effectively consider and measure the similarity between these entities. By employing cosine similarity, the model incorporates a quantitative assessment of the similarity, which aids in discerning both the resemblances and disparities within and across different classes.

In the following section, we would like to provide the details of the used backbone in the proposed method, named Shifted-Dilated Window-based Self-Attention (S-DWSA) Transformer.

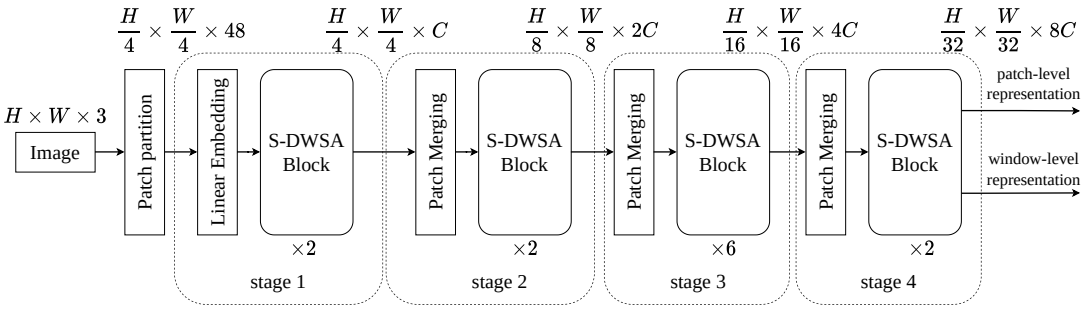


Figure 6.8: *The architecture of the S-DWSA Transformer.*

4.2 S-DWSA Transformer backbone

An overview of the S-DWSA Transformer architecture is presented in Figure 6.8. It first splits an input RGB image into non-overlapping patches by a patch-splitting module, like ViT. Each patch is treated as a "token" and its feature is set as a concatenation of the raw pixel RGB values. In our implementation, we use a patch size of 4×4 and thus the feature dimension of each patch is $4 \times 4 \times 3 = 48$. A linear embedding layer is applied to this raw-valued feature to project it to an arbitrary dimension (denoted as C).

Several S-DWSA Transformer blocks are applied to these patch tokens. The S-DWSA Transformer blocks maintain the number of tokens ($H/4 \times W/4$), and together with the linear embedding are referred to as "Stage 1".

To produce a hierarchical representation, the number of tokens is reduced by patch-merging layers as the network gets deeper. The first patch merging layer concatenates the features of each group of 2×2 neighboring patches and applies a linear layer on the 4C-dimensional concatenated features. S-DWSA Transformer blocks are applied afterward for feature transformation, with the resolution kept $\frac{H}{8} \times \frac{W}{8}$. This first block of patch merging and feature transformation is denoted as "Stage 2". The procedure is repeated twice, as "Stage 3" and "Stage 4", with output resolutions of $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, respectively.

4.2.1 S-DWSA Transformer block

The S-DWSA Transformer is established by substituting the standard multi-head self-attention (MSA) module within a Transformer block with a module inspired by the shifted windows technique derived from the Swin Transformer [306], while maintaining consistency in the remaining layers. As illustrated in Fig. 6.9, an S-DWSA Transformer block comprises a Shifted-Dilated Window-based MSA module, followed by a 2-layer

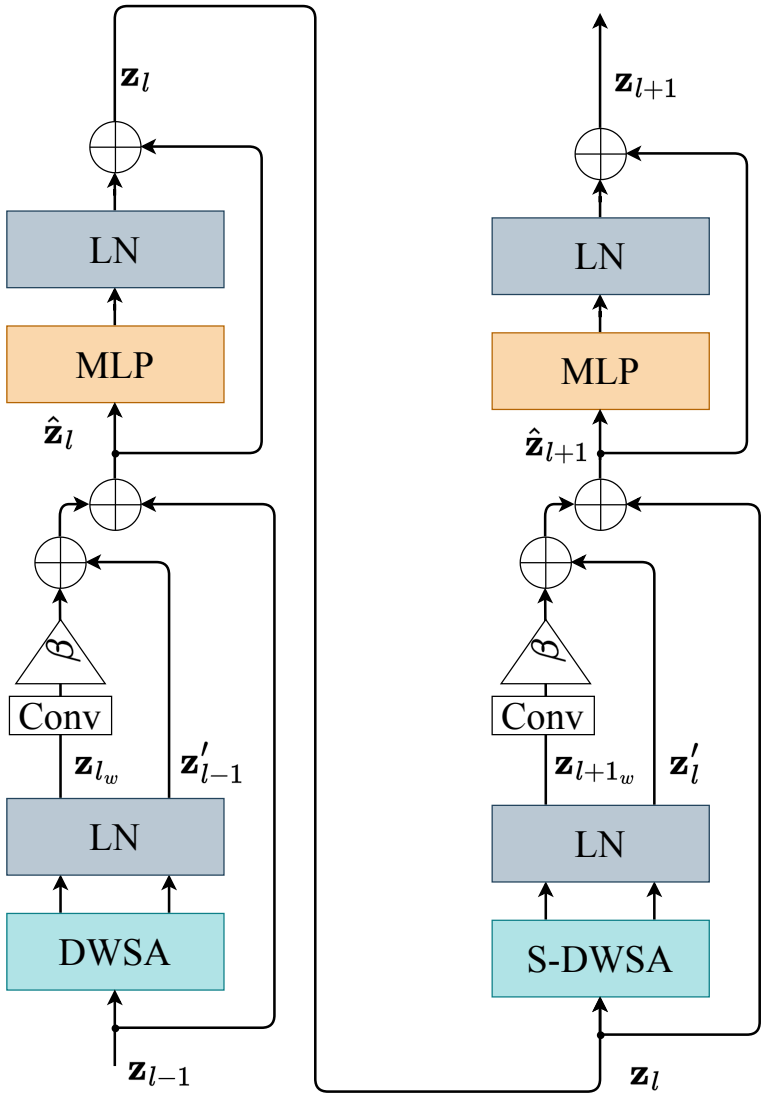


Figure 6.9: *Two successive S-DWSA Transformer Blocks.*

MLP incorporating GELU non-linear activation functions in between. Following each MSA module and each MLP, we apply a LayerNorm (LN) layer, and we incorporate a residual connection after each module.

It's notable that exploiting varied drop path rates within the residual connections in an asymmetric style has shown its benefits, particularly in addressing tasks like image classification and the application of Transformer architectures, as previously seen in works such as [306], [307]. Consequently, it becomes imperative to subject this aspect to thorough examination through an ablation study, which we will delve into further in Section 5.2.

Assuming that each window accommodates patches in a $P \times P$ arrangement, the

initial module utilizes a standard window partitioning strategy that commences from the top-left pixel. In this setup, the dilated space is established with a value of $ds = 1$ (as shown in Fig. 5.3). Subsequently, the subsequent module adopts a windowing configuration that deviates from the preceding layer's setup by shifting the windows by $(\lfloor \frac{P}{2} \rfloor, \lfloor \frac{P}{2} \rfloor)$ patches from the regularly partitioned windows. This shift in window partitioning also results in a corresponding shift in the dilated window. Therefore, consecutive computations of S-DWSA Transformer blocks are computed as follows:

$$[\mathbf{z}_{l_w}, \mathbf{z}'_{l-1}] = LN(DWSA(\mathbf{z}_{l-1})), \tag{6.13}$$

$$\hat{\mathbf{z}}_l = [\mathbf{z}'_{l-1} + \beta(Conv(\mathbf{z}_{l_w}))] + \mathbf{z}_{l-1}, \tag{6.14}$$

$$\mathbf{z}_l = LN(MLP(\hat{\mathbf{z}}_l)) + \hat{\mathbf{z}}_l, \tag{6.15}$$

$$[\mathbf{z}_{l+1_w}, \mathbf{z}'_l] = LN(S\text{-}DWSA(\mathbf{z}_l)), \tag{6.16}$$

$$\hat{\mathbf{z}}_{l+1} = [\mathbf{z}'_l + \beta(Conv(\mathbf{z}_{l+1_w}))] + \mathbf{z}_l, \tag{6.17}$$

$$\mathbf{z}_{l+1} = LN(MLP(\hat{\mathbf{z}}_{l+1})) + \hat{\mathbf{z}}_{l+1}, \tag{6.18}$$

where $\mathbf{z}_l$ and $\mathbf{z}_{l_w}$ denote the output features regarding patch-level attention and window-level attention at the layer l, β is the regular parameter that controls the impact of two attention elements, respectively; Conv denotes the convolution layer for shape consistency. DWSA and S-DWSA denote Dilated Window-based Self-Attention using regular and shifted window partitioning configurations, respectively. The shifted window partitioning approach introduces connections between neighboring non-overlapping windows in the previous layer and the relation between surrounding windows which is found to be effective in image classification. In the subsequent section, we will present the details of the Dilated Window-based Self-Attention (DWSA) module.

4.2.2 Shifted-Dilated Window-based Self-Attention

The standard Transformer architecture for image classification [308] both conduct global self-attention, where the relationships between a token and all other tokens are computed. The global computation leads to quadratic complexity with respect to the number of tokens, making it unsuitable for many vision problems requiring an immense set of tokens for dense prediction or to represent a high-resolution image. Inspired by the Shifted Window-based Self-Attention (S-WSA) of Swin Transformer [306], we propose a novel Shifted-Dilated Window-based Self-Attention (S-DWSA) to simultaneously compute self-attention within local attention windows and the attention based on surrounding windows (Fig. 5.3). The Dilated Window-based Self-Attention (DWSA) is formulated

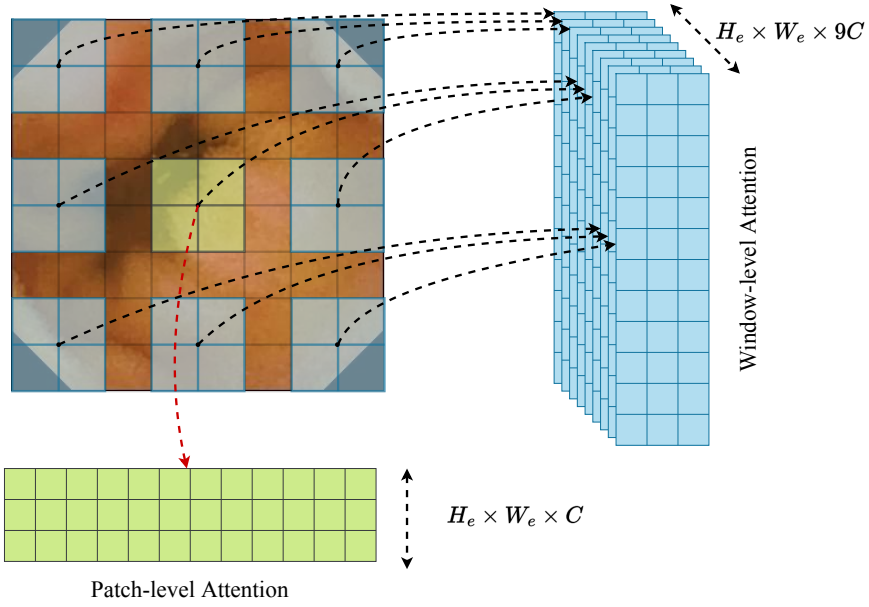


Figure 6.10: *An illustration of the Shifted-Dilated Window-based Self-Attention. In layer l, a regular window partitioning scheme is adopted, and self-attention is computed within each window (patch-level). On the right, the window-level attention is calculated based on 9 dilated surrounding windows. In layer $l+1$, the window partitioning is shifted, and the self-attention computation in the new windows crosses the boundaries of the previous windows in layer l, providing overlapping among them.*

as:

$$\mathbf{z} = Attention(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax(\frac{sim(\mathbf{Q},\mathbf{K}^\top)}{\sqrt{d}} + \mathbf{B})\mathbf{V}, \tag{6.19}$$

where $\mathbf{B}$ is the relative position bias term for each attention head; $\mathbf{Q},\mathbf{K},\mathbf{V}$ are the query, key, and value matrices; d is the query/key dimension. The relative position bias encodes relative spatial configurations of visual elements and is shown critical in a variety of visual tasks, especially for dense recognition tasks such as object detection.

However, in the original self-attention computation, the similarity terms of the pixel pairs are computed as a dot product of the query and key vectors. As mentioned in [309], when this approach is used in large visual models, the learned attention maps of some blocks and heads are frequently dominated by a few pixel pairs. To ease this issue, in the Dilated Window-based Self-Attention (DWSA) module, the attention logit of query $\mathbf{q}_i \in \mathbf{Q}$ and key $\mathbf{k}_j \in \mathbf{K}$ by a scaled soft-cosine similarity approach has been proposed as:

$$sim(\mathbf{q}_i,\mathbf{k}_j) = \frac{\sum_{m,n}^{N}(s_{mn}\mathbf{q}_{im}\mathbf{k}_{jn})}{\sqrt{\sum_{m,n}^{N} s_{mn}\mathbf{q}_{im}\mathbf{q}_{in}}\sqrt{\sum_{m,n}^{N} s_{mn}\mathbf{k}_{jm}\mathbf{k}_{jn}}}/t_s, \tag{6.20}$$

where $s_{mn} = \text{similarity}(feature_m, feature_n)$ and t_s is a learnable scalar, non-shared

across heads and layers. In this work, the Jaccard coefficient is considered as the similarity index. If there is no similarity between features ($s_{mn} = 1, s_{mn} = 0$ for $m \neq n$), the given equation is equivalent to the conventional scaled-cosine similarity formula. Unlike scaled-cosine similarity proposed in [309], scaled soft-cosine similarity can capture the underlying similarities, taking into account not just the direction of the feature vectors, but also their semantic magnitude. This enables a better comparison, allowing for accurate discrimination to handle inter-class similarity and intra-class variance in WCE image classification.

5 Experimental results

To assess the efficiency of the novel methodologies, comprehensive experiments have been undertaken. The initial set of experimental trials is aimed at studying the impact exerted by distinct parameters inherent to the formulated S-DWSA framework. Subsequently, a comparative analysis is executed between the aforementioned S-DWSA architecture and contemporary cutting-edge classification techniques. This comparative assessment is designed to discern the relative performance and capabilities of the proposed approach concerning established state-of-the-art methods within the domain of WCE classification.

5.1 Experimental setting

In this subsection, we will focus on the optimization process, provide a comprehensive overview of the dataset employed, and explore the selection of metrics for conducting meaningful performance comparisons.

Following the preceding section, the proposed methodology encompasses a dual-encoder configuration: an online encoder and a target encoder. Each of these encoders comprises a foundational backbone and a projector head, realized as a two-layer Multi-Layer Perceptron (MLP). A distinctive attribute of the online encoder is the inclusion of an auxiliary prediction head, also a two-layer MLP. This introduction of asymmetry between the two encoders delineates their respective roles.

5.1.1 Optimization

The training process of the online encoder is updated through gradient-based updates, while the target encoder is refined via a momentum-based moving average mechanism (EMA) during each iteration of the training process. The training process embraced by our models follows a two-stage routine. The initial phase encompasses what we

refer to as the ”pretext task,” which acts as the upstream pre-training stage. In this step, both the online and target branches facilitate the acquisition of intermediate representations through the use of unlabeled WCE images. The mechanism through which knowledge is transmitted from the online branch to the target branch embodies a transfer process, wherein updates propagated through the online branch intricately refine and exclusively enhance the target branch. This process stage is accomplished over 100 epochs, indicating a defined period during which this knowledge transfer phenomenon transpires. The optimization detail of the proposed network is shown in Algorithm 6.1.

Subsequently, in the second step, the comprehensive model training proceeds, exclusively centering its focus on the target model. The training process is geared towards the specific task of WCE image classification. Notably, the downstream tasks in this context are reliant on a more limited reservoir of labeled WCE images. A consistent learning rate of 0.001 and a stable weight decay rate of 0.05 are employed in the experimental sphere, yielding commendable performance. The optimization of hyperparameters is a critical endeavor in our methodology. Parameters such as the key queue size, the initial momentum value for the target encoder, the temperature τ, the influence factor β, and the drop path rates are all subject to hyperparameter tuning, ensuring an optimal configuration for the peak performance.

5.1.2 Dataset

Experiments were conducted using the Kvasir-Capsule dataset, as outlined in the work of Smedsrud et al. (2021) [10]. This dataset encompassed a total of 38,837 well-labeled WCE images representing eight distinct pathological conditions as shown in Fig. 6.11, alongside an additional 400,000 unlabeled data instances.

More precisely, the labeled WCE images were subjected to a randomized allocation process to facilitate a four-fold cross-validation procedure, ensuring equitable distribution across all pathology categories. Notably, the distribution of data within each pathological class remained uniform after the random partitioning. To diversify the available data, a set of augmentations, inclusive of horizontal and vertical flips, as well as rotations (at angles of 0°, 90°, 180°, and 270°), were systematically employed, resulting in an eight-fold expansion of the data volume. The subsequent visual representation presents the distribution of labeled images associated with each distinct class.

Algorithm 6.1: Pseudo code of DWSA-SSL in a PyTorch-like style.

```
# encoder:  S-DWSA transformer-based encoder
# proj:  projector
# pred:  predictor
# odpr:  online drop path rate
# tdpr:  target drop path rate
# m:  momentum coefficient
# t:  temperature coefficient
# queue1, queue2:  two queues for storing negative samples
f_online = lambda x:  pred(proj(encoder(x, drop_path_rate=odpr)))
f_target = lambda x:  proj(encoder(x, drop_path_rate=tdpr))
for v1, v2 in loader:  # load two views
   q1', q2', q1'_w, q2'_w = f_online(v1), f_online(v2) # queries:
    N×C
   k1", k2", k1"_w, k2"_w = f_target(v1), f_target(v2) # keys:
    N×C
   # symmetric loss
   loss = window_contrastive_loss(q1', k2", q1'_w, k2"_w, queue2)
    + window_contrastive_loss(q2', k1", q2'_w, k1"_w, queue1)
   loss.backward()
   update(f_online)# optimizer update:  f_online
   f_target = m * f_target + (1.  - m) * f_online # momentum
    update:  f_target
   update(m) # update momentum coefficient
def contrastive_loss(q, k, q_w, k_w, queue):
   k_f = k/cos(q_w,k_w)
   # positive logits:  N×1
   l_pos = torch.einsum('nc,nc->n', [q,
    k_f.detach()]).unsqueeze(-1)
   # negative logits:  N×K
   l_neg = torch.einsum('nc,ck->nk', [q, queue.clone().detach()])
   # logits:  N×(1+K)
   logits = torch.cat([l_pos, l_neg], dim=1)
   # labels:  positive key indicators
   labels = torch.zeros(N)
   loss = F.cross_entropy(logits / t, labels)
   # update queue
   enqueue(queue, k)
   dequeue(queue)
return loss
```

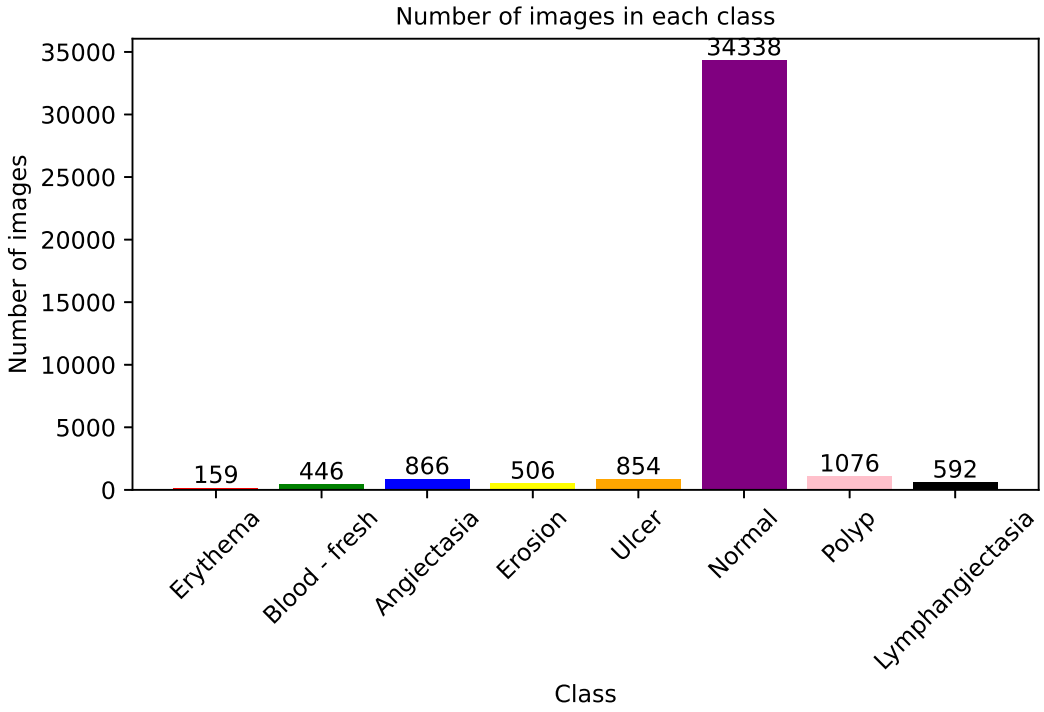


Figure 6.11: *The number of images in the various Kvasir-Capsule labeled image classes.*

5.1.3 Evaluation metrics

Five assessment metrics were employed to evaluate the effectiveness of the classification process. These metrics encompass the overall accuracy of the classification, the precision which quantifies the ratio of correctly predicted positive observations to the total predicted positives, the recall which measures the ratio of correctly predicted positive observations to the total actual positives, the F1-score that harmonizes precision and recall, and lastly, the Matthews correlation coefficient (MCC) that encapsulates the quality of binary classification predictions. Furthermore, in the context of a comprehensive evaluation, computations were conducted to determine the count of parameters involved in the process as well as the number of Floating-Point Operations Per Second (FLOPS), thereby facilitating a comprehensive benchmarking analysis.

In the subsequent section, these specified metrics will play a pivotal role in the establishment of an ablation study. This study is designed to meticulously assess and understand the performance and impact of each discrete constituent comprising the proposed methodology. Through this examination, we seek to gain comprehensive insights into the contribution and effectiveness of every element within the proposed

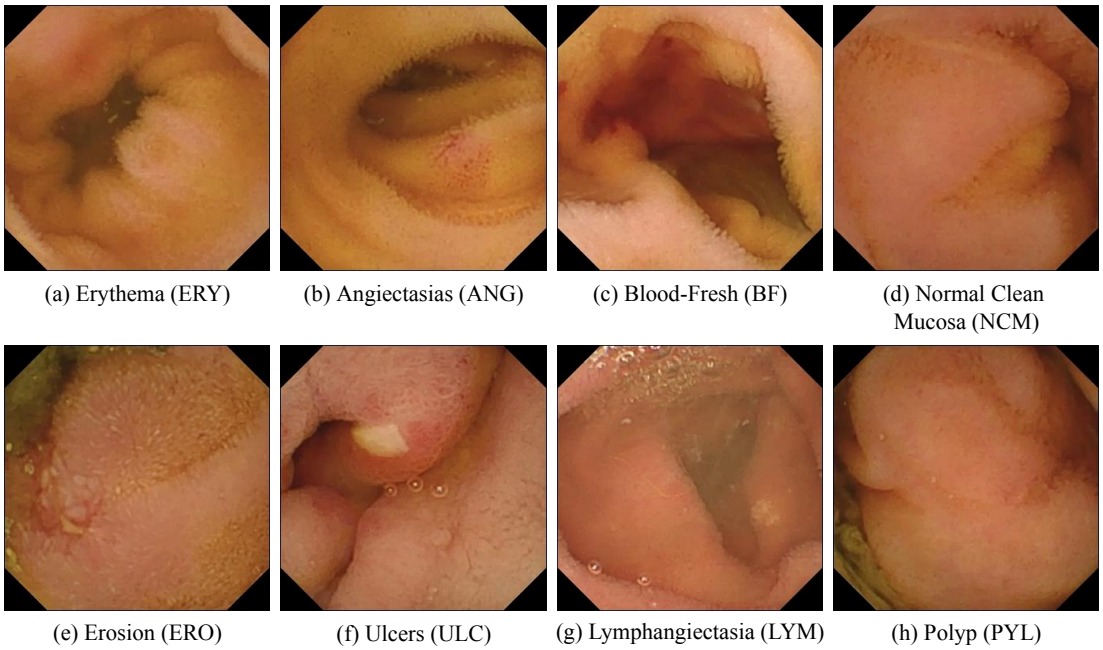


Figure 6.12: *Eight pathological classes in the Kvasir-Capsule dataset [10].*

approach. By systematically dissecting and evaluating each component, we aim to discern their individual significance and the collective impact they exert on the overall system performance. This endeavor will give a better understanding of the inner mechanism and will guide us in optimizing its constituent parts for enhanced outcomes.

5.2 Ablation study

In this dedicated section, a meticulous process of ablation is embarked upon, strategically focusing on pivotal design elements inherent to the proposed network architecture.

Firstly, a comprehensive analysis is dedicated to the hyper-parameters employed within the proposed network. This investigation aims to focus on the effects and implications of the chosen hyper-parameters on the network's performance and overall functionality. Through meticulous examination, we seek to uncover the relationships between the various hyper-parameters and their impact on the network's behavior. This analytical effort ultimately contributes to a deeper understanding of the network's intricacies, allowing for informed adjustments and optimizations that can potentially enhance its performance.

Subsequently, we turn our attention to assessing the performance disparities between the Shifted-Dilated Window-based Self-Attention mechanism and the initial Window-based Self-Attention design. Additionally, an in-depth analysis is conducted on the

scale-soft-cosine attention approach. Through rigorous and systematic analysis, our objective is to unravel the efficacy and consequential influence exerted by these distinct attention mechanisms upon the overall performance of the network. This experiment enables us to extract valuable insights that can aid in shaping a well-informed decision-making process, especially in the context of selecting the most suitable attention mechanism for realizing optimal network outcomes.

5.2.1 Ablation on hyper-parameters

Within each ablation experiment, a singular hyper-parameter is subjected to variation, while the remaining hyper-parameters are maintained at their default values. This approach allows for an isolated assessment of the impact that individual hyper-parameters exert on the system's behavior and performance. By systematically altering one hyper-parameter at a time and observing resultant changes, we gain insights into the intricate interplay between these parameters and the overall system dynamics. This methodological rigor ensures a comprehensive exploration of the hyper-parameter landscape, facilitating a nuanced understanding of their individual contributions.

Firstly, the employment of asymmetric drop path rates has proven to be advantageous. Drop path has emerged as an effective regularization technique in the context of supervised representation learning, particularly when dealing with image classification tasks and Transformer architectures. In our analysis, we also investigate the impact of this regularization strategy, as outlined in Table 6.1.

Table 6.1: *Ablation study on the drop path rates of online and target encoders, while holding other parameters at default including queue size* $K = 2048$*, temperature* $\tau = 0.1$*, momentum* $\alpha = 0.99$*, regular parameter* $\beta = 0.75$

Online dpr	Target dpr	Top-1	Precision (%)		Recall (%)		F1-Score (%)		MCC (%)
		Accuracy (%)	Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	
0.05	0.00	97.52	71.51	**98.89**	66.53	87.89	68.78	**98.88**	83.47
0.10	0.00	97.78	**71.65**	97.62	**66.74**	**87.98**	68.93	97.53	**85.36**
0.20	0.00	**97.87**	69.53	97.70	65.62	85.83	**69.80**	98.82	85.35
0.10	0.10	95.51	69.72	96.64	66.66	87.84	65.92	92.69	83.89

Increasing the drop path regularization rate from 0.05 to 0.1 for the online encoder has proven to be a strategic adjustment that brings about positive outcomes in the domain of representation learning. This favorable impact can be attributed to its capacity to potentially mitigate the prevailing issue of overfitting. By enhancing the regularization rate, the model demonstrates a greater ability to generalize and capture underlying patterns in the data, resulting in more robust and meaningful representations. However, when the same drop path regularization is introduced to the target encoder, a

contrasting and rather unexpected trend emerges. This adjustment leads to a discernible reduction of not only 2.36% in the top-1 accuracy score, marking a significant decline from 97.87% to 95.51% but also other comparative metrics. This noteworthy decrease in accuracy underscores a clear and detrimental impact on the performance of the target encoder. The pronounced drop in accuracy suggests that the higher regularization rate of 0.1 might be inducing excessive suppression of information within the target encoder. Instead of enhancing generalization, the increased regularization appears to hinder the encoder's ability to capture intricate details and fine-grained features present in the data. As a result, the model's capacity to discriminate and classify accurately is compromised, leading to a substantial degradation in overall performance.

Next, the temperature hyperparameter τ embedded within the loss function, as denoted by Equation (6.12), governs the degree to which the model prioritizes the negative samples from unlabeled data, thereby optimizing the discrimination of features. Our experimental investigations encompassed the evaluation of network performance across different values of τ. The corresponding comparative metrics are reported in Table 6.2.

Table 6.2: *Ablation study on the temperature τ, while holding other parameters at default including queue size $K = 2048$, momentum $\alpha = 0.99$, regular parameter $\beta = 0.75$ and online encoder drop rate of 0.1*

Temperature	Top-1	Precision (%)		Recall (%)		F1-Score (%)		MCC (%)
τ	Accuracy (%)	Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	
0.05	95.67	69.77	94.71	64.72	86.82	66.56	95.52	83.55
0.10	**97.78**	**71.65**	97.62	**66.74**	**87.98**	**68.93**	97.53	85.36
0.20	97.64	71.25	97.58	65.72	86.91	67.63	97.83	84.56
0.50	97.60	70.56	**97.71**	65.53	86.89	67.73	**98.50**	**86.57**

The results unveil a discernible pattern: as τ is varied within the range of 0.05 to 0.5, the performance of our method exhibits an initial increase followed by a subsequent decrease. It becomes evident that an extreme case of disregarding the feature distribution of unlabeled data ($\tau = 0.05$) fails to yield satisfactory classification performance. Conversely, a network trained with excessive emphasis on negative samples of unlabeled data ($\tau = 0.5$) is prone to deteriorating performance. This occurs due to the potential misclassification of truly pseudo-labeled samples into distinct classes, resulting in error accumulation.

The appropriately selecting the value of $\tau = 0.1$ effectively enhances the accuracy of deeply learned features, consequently elevating the overall classification performance. This empirical observation underscores the pivotal role that the temperature hyperparameter plays in striking a balance between leveraging negative samples for discrimination

and avoiding the pitfalls of overemphasis and misclassification.

Lastly, an analysis is conducted to ascertain the effectiveness of the queue size in the attention product. This analysis aims to provide insights into how the queue size influences the performance of the attention mechanism. By investigating the impact of varying queue sizes, we aim to uncover the optimal configuration that maximizes the mapping performance of the attention mechanism, thereby contributing to enhanced model outcomes.

Table 6.3: *Ablation study on the queue size, while holding other parameters at default including temperature $\tau = 0.1$, momentum $\alpha = 0.99$, regular parameter $\beta = 0.75$ and online encoder drop rate of 0.1.*

Queue Size	Top-1	Precision (%)		Recall (%)		F1-Score (%)		MCC (%)
K	Accuracy (%)	Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	
1024	95.83	68.59	95.52	64.70	85.70	65.82	96.76	**87.37**
2048	**97.78**	**71.65**	97.62	66.74	**87.98**	**68.93**	97.53	85.36
4096	96.72	67.70	98.57	**67.61**	86.98	68.84	**97.84**	80.46
8192	96.82	67.73	**99.51**	65.59	82.52	68.58	94.85	83.41

As we can see in Table 6.3, a larger queue size allows the model to store a more diverse and extensive set of features, thereby enhancing its capacity to recognize intricate patterns within the data. This, in turn, contributes to better generalization and higher accuracy on unseen data. However, over-increasing queue size might come at the cost of the model's overall performance, as it might not be able to learn and adapt to the underlying data distribution effectively.

5.2.2 Ablation on scale-soft-cosine attention

Table 6.4 provides a comprehensive evaluation of the performance impact derived from applying the proposed scale-soft-cosine attention methodologies within the suggested architectural framework. The experimental results showcase notable enhancements in accuracy across different attention logits. Specifically, the observed improvements are quantified at approximately +0.2%, and more than +0.3% of all metrics for the respective logits categories, thereby highlighting the pronounced benefits of these techniques.

A pivotal observation is the stabilizing influence exerted by the scale-cosine attention during the training process. As demonstrated in Fig. 6.13, the activation values for the S-DWSA Transformer at deeper layers exhibit a propensity for explosive growth in the case of the large model size. In contrast, the newly introduced version of the model demonstrates significantly milder behavior in this regard. This holds crucial significance, as the original attention multiplication mechanism leads to divergence

Table 6.4: *Influence of the scale-soft-cosine attention logits in the proposed architecture on the classification performance.*

Logits	Top-1 Accuracy (%)	Precision (%)		Recall (%)		F1-Score (%)		MCC (%)
		Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	
Scale-dot	98.27	70.91	97.79	65.36	86.55	67.59	96.70	84.37
Scale-cosine	**98.46**	71.10	**98.44**	66.19	87.21	68.12	96.71	84.81
Scale-soft-cosine	97.78	**71.65**	97.62	**66.74**	**87.98**	**68.93**	**97.53**	**85.36**

during self-supervised pre-training for expansive models, while the scale-soft-cosine attention effectively mitigates this issue, resulting in stable training progress.

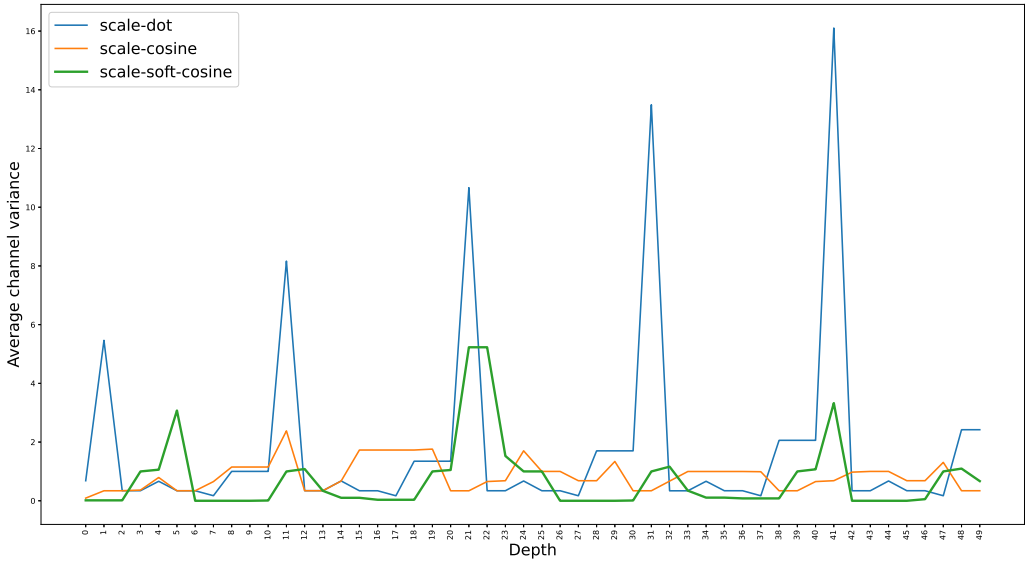


Figure 6.13: *The Signal Propagation Plot for various attention logits.*

Moreover, in comparison to the scale-cosine attention approach, the proposed soft-cosine attention manifests a more pronounced distribution of activation. This attribute facilitates superior propagation of activation signals across the neural network's layers, particularly during the initial stages of training (initialization). This dynamic aids in establishing a solid foundation for the training process and enhances the potential for efficient information transfer across the network. Therefore, we can confidently conclude that the utilization of the scale-soft-cosine attention mechanism yields tangible benefits for our proposed architecture. Therefore, in the subsequent subsection, we embark on an ablation study of the S-DWSA Transformer integrated with the scale-soft-cosine attention mechanism. This investigation is undertaken to highlight the superior performance and efficacy of the proposed method in comparison to other SSL architectures.

5.2.3 Ablation on the Shifted-Dilated Window-based Self-Attention

The outcomes of ablating the shifted window approach across the three other SSL architectures are reported in Table 6.5. Notably, the DWSA-SSL Transformer model demonstrates superior performance compared to other methods. Specifically, the improvements are quantified at +1.17%, +1.1%, and +0.92% top-1 accuracy for the respective methods, thereby highlighting the noticeable benefits of the proposed method.

Table 6.5: *Comparison of different SSL architectures.*

Method	Params (M)	FLOPs (G)	Top-1 Accuracy (%)	Precision (%)		Recall (%)		F1-Score (%)		MCC (%)
				Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	
MOCO v2	45	4.9	96.61	70.89	**97.76**	63.71	**88.95**	**76.71**	93.67	**86.90**
BYOL	42	4.6	96.68	66.86	95.68	58.69	86.52	61.54	95.51	76.79
MOBY	29	4.5	96.86	46.91	92.74	50.61	76.50	48.88	94.70	73.50
DWSA-SSL	31	4.6	**97.78**	**71.65**	97.62	**66.74**	87.98	68.93	**97.53**	85.36

These results notably underscore the efficacy of incorporating shifted windows as a mechanism for establishing interconnections among windows across preceding layers. The substantial performance gains across diverse tasks affirm the advantage of this approach, suggesting that the utilization of shifted windows contributes to enhanced modeling capabilities and more effective information propagation throughout the network. This phenomenon can be attributed to the fact that the encoder concentrates its processing on crucial regions, precisely determined by the Shifted-Dilated Window-based Self-Attention (S-DWSA) modules. This focused attention enables the extraction of pertinent features, thereby facilitating precise and robust classification of abnormalities.

Furthermore, it's important to note that the incorporation of window-level attention confers additional benefits to the architecture, resulting in a noteworthy 0.92% improvement in accuracy. This is owing to the fact that the window-level attention map is capable of encapsulating valuable texture information from the neighboring areas. The integration of such texture details proves advantageous, subsequently leading to enhanced classification performance. This interplay between patch-level and window-level attention mechanisms thus demonstrates their synergistic effect in achieving improved classification outcomes.

To further analyze the effectiveness of the window-level attention feature, we turn our attention to the hyper-parameter β. The purpose of this analysis is to delineate the optimal combination that maximizes the performance benefits achieved through the synergy of these attention mechanisms.

As discernible from Table 6.6, an interesting pattern becomes evident concerning the hyper-parameter β. Specifically, when β assumes a value of 0.75, the method attains

Table 6.6: *Influence of the window-level attention in the proposed architecture on the classification performance.*

Beta	Top-1 Accuracy (%)	Precision (%)		Recall (%)		F1-Score (%)		MCC (%)
		Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	
0.50	95.70	68.64	**98.53**	65.94	86.87	68.76	94.98	84.57
0.75	**97.78**	**71.65**	97.62	**66.74**	**87.98**	**68.93**	**97.53**	**85.36**
1.00	95.83	70.53	95.76	63.57	85.62	67.98	94.54	82.64
1.50	92.64	67.71	93.59	64.77	84.63	67.88	93.88	82.64
2.00	92.83	65.91	92.50	63.67	81.89	61.71	91.68	80.51

its peak performance. This observation underscores the pivotal role played by this hyper-parameter in influencing the method's efficacy. A value of β at this level effectively capitalizes on the salient information originating from the adjacent regions, resulting in heightened performance outcomes. However, a subsequent increase in β triggers an influx of redundant information, ultimately culminating in misclassification events. This empirical insight underscores the delicate balance that β strikes between leveraging surrounding context and avoiding the perils of information redundancy, consequently delineating the optimal configuration for achieving peak method performance. Subsequent to conducting the ablation study, the optimal parameters that yield the most optimal performance characteristics are selected for the purpose of comparing against the current state-of-the-art benchmarks.

5.3 Comparison with state-of-the-art methods

In this section, we looked at how well our method works by comparing it to eleven other related approaches in classifying images from WCE. We also explained the other methods we compared against in Section 2. We tried out these methods on our own dataset and showed the results for comparison.

Table 6.7: *Performance comparison with state-of-the-art methods.*

Method	Top-1 Accuracy (%)	Precision (%)		Recall (%)		F1-Score (%)		MCC (%)
		Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	Macro-avg	Weighted-avg	
Sekuboyina et al.[292]	91.64	27.58	84.74	25.68	91.60	26.70	87.50	46.67
Sindhu et al.[293]	88.58	23.62	81.79	17.50	88.70	18.53	83.80	47.41
Sharif et al.[294]	93.99	25.75	89.94	33.78	93.61	28.92	91.58	74.30
Cao et al.[295]	86.54	30.90	78.57	31.93	86.75	28.73	81.61	73.48
Lan et al.[226]	96.51	63.59	95.80	61.72	**96.96**	66.94	94.79	45.07
Sadasivan et al.[222]	89.55	30.75	81.98	36.75	89.76	32.55	85.83	57.57
Xing et al.[296]	93.83	53.61	90.84	38.75	93.79	40.76	90.85	58.69
Xing et al.[227]	96.66	66.74	95.61	58.92	96.52	61.75	95.80	79.99
Guo et al.[297]	96.93	46.75	92.66	50.70	96.74	48.50	94.69	40.76
Guo et al.[298]	94.97	35.54	88.83	40.69	94.62	37.91	91.98	61.41
Zhao et al.[299]	90.89	25.51	81.82	27.91	90.78	26.75	96.69	76.33
DWSA-SSL	**97.81**	**71.53**	**97.50**	**66.79**	87.58	**68.85**	**97.58**	**85.36**

Our comprehensive evaluation is presented through two main channels: Firstly, table 6.7 exhibits metrics used to assess the overall performance in comparison to state-of-the-art techniques. The results are unequivocal: Our approach surpasses established methodologies in the domain of WCE image classification. Notably, the work by Xing et al. (2020) [227] emerges as the closest competitor across all evaluation criteria. Particularly remarkable is the significant enhancement our method showcases over this benchmark, achieving an improvement of 2-3% across all evaluated metrics. This accentuates the capability of our method to discern crucial attributes within WCE images, even when confronted with unlabelled data.

Table 6.8: *Model complexity comparison with state-of-the-art methods.*

Method	**Params (M)**	**FLOPs (G)**	**Throughput images / s**
Sekuboyina et al.[292]	7.6	1.1	2665.2
Sindhu et al.[293]	3	1	2691.6
Sharif et al.[294]	220	35	94.4
Cao et al.[295]	132	7.63	832.8
Lan et al.[226]	138	15.5	410.8
Sadasivan et al.[222]	143	19.67	143.6
Xing et al.[296]	102	9	1261.6
Xing et al.[227]	86	18.25	343.6
Gou et al.[297]	102	9	1262
Gou et al.[298]	46	5.2	898.8
Zhao et al.[299]	86	16.85	332.8
DWSA-SSL	31	4.6	1060.8

Secondly, our exploration extends to the realm of model complexity. Our model comprises 31 million parameters, significantly fewer than other deep learning-based architectures like Sharif et al. [294] with about 220 million parameters, and VGG16-based models (Cao et al. [295], Lan et al. [226]) with around 130 million parameters. Despite the fact that our attention-based model integrates a more extensive array of parameters compared to the conventional machine learning approaches, the computational overhead associated with this augmented complexity remains well-contained and manageable. It is noteworthy that the heightened intricacy within our model is counterbalanced by its remarkable performance, which notably outshines the results achieved by traditional machine learning methods. This substantiates the appropriateness of the investment in terms of model complexity, considering the substantial enhancement in predictive

capability and overall effectiveness.

Turning to computational efficiency, due to our meticulous optimization of Floating Point Operations (FLOPs), the throughput of our approach has undergone a noteworthy augmentation. This enhancement is particularly remarkable as it is a direct result of our method's efficient utilization of computational resources, leading to substantial gains in processing speed and overall efficiency. More precisely, our model achieves a diagnostic speed of approximately 1000 images/ sec. For a complete WCE video containing 55,000 images, our method requires around 55 seconds for analysis, using a batch size of 32. This real-time diagnostic capability, alongside its strong accuracy, underscores the promising potential of our approach in real-world clinical applications.

6 Conclusion

In this chapter, we introduce a pioneering self-supervised learning paradigm, harnessing a revolutionary Dilated Window Self-Attention Self-Supervised Learning (DWSA-SSL) as its foundational architectural framework. Distinguished from the Vision Transformer (ViT), the novel S-DWSA Transformer backbone grants us the capacity to not only appraise the acquired representations but also subject them to a comprehensive evaluation in subsequent tasks, such as the classification of pathologies within the context of WCE images.

A distinctive contribution of our method is the application of window-level attention, a concept that guides the model's focus not only toward the lesion region but also toward the pertinent neighboring areas. By facilitating this expanded focus, the model becomes proficient in incorporating contextual cues, thereby empowering it to aptly classify the diverse pathologies found within WCE images. As evidenced by our findings, the performance of the proposed method was significantly overcome with, or experiences only slight deviation from, the results achieved by supervised methods. This nuanced outcome signifies a potential avenue for refinement within the domain of self-supervised learning with Transformer architectures. Our optimism is that these outcomes will catalyze advancements in self-supervised learning methods designed specifically for Transformer architectures, ushering in a new era of enhanced performance and understanding in the realm of medical image analysis.

CHAPTER 7

Conclusion and perspective

Chapter content

1 Summary and conclusion 154

2 Future work and perspectives 156

This chapter provides a summary of the work presented and the conclusions drawn from this work. It lists the contributions to knowledge already achieved in this research and provides directions for future work.

1 Summary and conclusion

In this thesis, we have investigated some issues related to image classification in WCE imaging. In particular, we have focused on an important factor, namely image quality, affecting the image classification performance.

Chapter 2: In this chapter, we introduced and discussed the key concepts of our proposed methods, which encompass aspects like No-Reference Image Quality Assessment (NR-IQA) metrics, image quality enhancement techniques, and classification methods. Furthermore, we delved into the existing research works that have been instrumental in shaping and inspiring our research journey throughout the course of my thesis. This chapter provides a foundational understanding of the concepts and influences that underpin our research endeavors.

Our initial task, as detailed in Chapter 3, revolves around the development of a dataset, a fundamental component crucial for the effective implementation of learning-based enhancement techniques. This dataset serves as the basis upon which our enhancement method relies for effective learning and improvement. To address the dearth of dedicated quality assessment datasets tailored for Wireless Capsule Endoscopy (WCE) videos, we proposed the creation of the Quality-Oriented Database for Video Capsule Endoscopy (QVCED), a dataset specifically designed to evaluate the quality of WCE videos. It comprises a diverse array of scenarios, encompassing various pathologies and multiple types of distortions at varying levels, thereby emphasizing realism. The dataset creation process involves two stages: the careful selection of reference videos meeting quality criteria from the Kvasir-Capsule dataset and subsequently subjecting these reference videos to a degradation process, simulating distortions to generate a comprehensive dataset with diverse video quality variations. This work underscores the dataset's significance, showcasing its diversity and broad applicability through a series of experiments across different modalities.

Chapter 4 focused on addressing uneven illumination (UI) issues in laparoscopy and WCE images. We introduce a novel UI assessment metric, namely Illumination Histogram Equalization Difference (IHED), based on the impact of contrast enhancement on image background illuminance. By quantifying how histogram equalization affects the spatial distribution of background illuminance, we can effectively evaluate and measure uneven illumination levels. To enhance sensitivity in scenarios with subtle variations, we propose an index that divides the standard deviation of the brightness channel by the mean intensity. Experiments on medical images validate our method's superior performance compared to existing IQA methods. This UI assessment metric plays a crucial role in our image enhancement methodology, enabling precise identification of areas affected by uneven illumination and ultimately improving image quality and diagnostic capabilities in different medical imaging modalities.

The creation of our specialized dataset for video quality assessment in the domain of Wireless Capsule Endoscopy (WCE) and the proposed quality assessment metric are foundational steps in the development of our image enhancement method. In our third research work, Chapter 5, we introduced the Triplet Clustering Fusion Autoencoder for Quality Enhancement of WCE images (TCFA) method for enhancing WCE image quality, drawing inspiration from attention-based advantages. TCFA employs a dual-branch autoencoder hierarchical network comprising the Distorted Image Projector (DIP), Distorted Image Encoder (DIE), Distorted Image Decoder (DID), Distortion Level Encoder (DLE), and Variational Cross-Attention Module (VCAM). The method

is extensively evaluated on images with varying distortion levels, demonstrating its superior performance compared to existing methods. These evaluations rigorously assess the method's effectiveness in addressing various image distortions, consistently showcasing its enhanced capabilities in restoring and enhancing image quality.

In our final research contribution, we expand upon the ideas introduced in Chapter 6 by delving into Self-Supervised Learning (SSL) for image representation. Our SSL approach outperforms standard supervised methods across seven downstream tasks and offers a promising alternative to traditional Convolutional Neural Networks (CNNs) with its Transformer-based backbone. The SSL network comprises two encoders, online and target, with an asymmetrical structure. During training, the online encoder updates with gradients, while the target encoder evolves through a moving average of the online encoder. Within each encoder, a hybrid attention mechanism within the Transformer-based backbone captures essential contextual information. This design enhances the SSL network's robustness and discriminative abilities, especially in classifying WCE images.

2 Future work and perspectives

Apart from some promising results, the proposed methods have still some limitations which need to be improved. They are listed below:

i. The effectiveness of the learning-based methods proposed in this thesis relies significantly on the collection of standard image datasets used for building the training set. This is because the training set might be skewed towards standard images. Typically, a larger dataset of standard images is essential to create a high-quality training set. Consequently, considerable memory capacity is often needed for this purpose.

ii. It's important to note that our work does not encompass all possible types of distortion that can occur in WCE images, and real-life scenarios may introduce additional distortions like specular reflection. This is particularly noteworthy because our IQA metrics rely on the presence of background illuminance. The presence of specular reflection in WCE images has the potential to impact the performance of these IQA metrics. Specular reflection, due to its unique characteristics, can introduce variations in the illuminance, potentially affecting the accuracy of the quality assessment process. Therefore, the presence of such distortions should be considered when applying IQA metrics to WCE images in real-world situations.

iii. Both the image enhancement method and the classification method heavily rely on the attention mechanism. It's important to note that training these methods with attention mechanisms can be computationally intensive and time-consuming. Additionally, their implementation demands substantial computational hardware resources to achieve optimal results. Therefore, the efficient utilization of attention mechanisms in these methods necessitates robust hardware support and patience during the training process.

As perspectives, some issues will be considered in the future. We can list some important tasks as follows:

i. To address the reliance on standard image datasets, future research should focus on the creation of more diverse and comprehensive training datasets that encompass a wider range of image variations and distortions. This could involve collecting real-world data and incorporating different levels of distortion, including less common artifacts like specular reflection, etc. Moreover, the previously defined hybrid distribution could be employed to accurately simulate uneven illumination, enhancing the realism of our distortion models.

ii. Developing IQA metrics that are more resilient to various types of distortion, including specular reflection, is crucial. Future research can focus on designing metrics that are less sensitive to background illuminance and better equipped to handle complex illumination scenarios.

iii. Further research into algorithmic efficiency, such as designing attention mechanisms that require fewer computations, can enhance the speed and efficiency of attention-based models.

> *"Any intelligent fool can make things bigger, more complex, and more violent. It takes a touch of genius and a lot of courage to move in the opposite direction."*
>
> E. F. Schumacher

Bibliography

[1] M. Arnold, C. C. Abnet, R. E. Neale, *et al.*, "Global burden of 5 major types of gastrointestinal cancer," *Gastroenterology*, vol. 159, no. 1, pp. 335–349, 2020 *Cited on page 2.*

[2] S. E. Roberts, D. G. Samuel, J. G. Williams, *et al.*, "Survey of digestive health across europe," *Part one: The burden of gastrointestinal diseases and the organisation and delivery of gastroenterology services across Europe. Report for United European Gastroenterology*, 2014 *Cited on page 2.*

[3] M. Owais, M. Arsalan, T. Mahmood, J. K. Kang, and K. R. Park, "Automated diagnosis of various gastrointestinal lesions using a deep learning–based classification and retrieval framework with a large endoscopic database: Model development and validation," *Journal of medical Internet research*, vol. 22, no. 11, e18563, 2020 *Cited on page 2.*

[4] M. Hmoud Al-Adhaileh, E. Mohammed Senan, W. Alsaade, *et al.*, "Deep learning algorithms for detection and classification of gastrointestinal diseases," *Complexity*, pp. 1–12, 2021 *Cited on page 2.*

[5] Y.-C. Wang, J. Pan, B. Jiang, *et al.*, "Direct visualization of drug behaviors in the upper gi tract via magnetically controlled capsule endoscopy," *VideoGIE*, vol. 6, no. 7, pp. 333–338, 2021 *Cited on page 2.*

[6] A. S. Ashour, N. Dey, W. S. Mohamed, *et al.*, "Colored video analysis in wireless capsule endoscopy: A survey of state-of-the-art," *Current medical imaging*, vol. 16, no. 9, pp. 1074–1084, 2020 *Cited on pages 2, 3.*

[7] I. M. Mehedi, K. P. Rao, F. M. Alotaibi, and H. M. Alkanfery, "Intelligent wireless capsule endoscopy for the diagnosis of gastrointestinal diseases," *Diagnostics*, vol. 13, no. 8, p. 1445, 2023 *Cited on pages 2–4.*

[8] O. S. Lin, "Sedation for routine gastrointestinal endoscopic procedures: A review on efficacy, safety, efficiency, cost and satisfaction," *Intestinal research*, vol. 15, no. 4, pp. 456–466, 2017 *Cited on page 3.*

[9] A. Koulaouzidis, K. Dabos, M. Philipper, E. Toth, and M. Keuchel, "How should we do colon capsule endoscopy reading: A practical guide," *Therapeutic Advances in Gastrointestinal Endoscopy*, vol. 14, p. 26 317 745 211 001 983, 2021 *Cited on page 3.*

[10] P. H. Smedsrud, V. Thambawita, S. A. Hicks, *et al.*, "Kvasir-Capsule, a video capsule endoscopy dataset," *Scientific Data*, vol. 8, no. 1, p. 142, 2021 *Cited on pages 5, 8, 9, 56, 57, 141, 144.*

[11] Y. Gao, W. Lu, X. Si, and Y. Lan, "Deep model-based semi-supervised learning way for outlier detection in wireless capsule endoscopy images," *IEEE Access*, vol. 8, pp. 81 621–81 632, 2020 *Cited on page 4.*

[12] S. H. Kim and Y. J. Lim, "Artificial intelligence in capsule endoscopy: A practical guide to its past and future challenges," *Diagnostics*, vol. 11, no. 9, p. 1722, 2021 *Cited on page 4.*

[13] F. Deeba, S. K. Mohammed, F. M. Bui, and K. A. Wahid, "A saliency-based unsupervised method for angiectasia detection in endoscopic video frames," *Journal of Medical and Biological Engineering*, vol. 38, pp. 325–335, 2018 *Cited on page 5.*

[14] T. Gan, S. Liu, J. Yang, B. Zeng, and L. Yang, "A pilot trial of convolution neural network for automatic retention-monitoring of capsule endoscopes in the stomach and duodenal bulb," *Scientific Reports*, vol. 10, no. 1, p. 4103, 2020 *Cited on page 5.*

[15] A. Mohammed, I. Farup, M. Pedersen, Ø. Hovde, and S. Yildirim Yayilgan, "Stochastic capsule endoscopy image enhancement," *Journal of Imaging*, vol. 4, no. 6, p. 75, 2018 *Cited on page 6.*

[16] V. B. S. Prasath, D. N. Thanh, L. T. Thanh, N. San, and S. Dvoenko, "Human visual system consistent model for wireless capsule endoscopy image enhancement and applications," *Pattern Recognition and Image Analysis*, vol. 30, pp. 280–287, 2020 *Cited on pages 6, 46, 56, 96, 109.*

[17] J. Modersitzki, "Numerical methods for image registration," in *Numerical Mathematics and Scientific Computation*, OUP Oxford, 2003 *Cited on page 15.*

[18] Y. Chen and J. Lee, "A review of machine-vision-based analysis of wireless capsule endoscopy video," *Diagnostic and Therapeutic Endoscopy*, vol. 2012, 2012 *Cited on pages 16, 56, 92.*

[19] S. Zou, M. Long, X. Wang, X. Xie, G. Li, and Z. Wang, "A CNN-based blind denoising method for endoscopic images," in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2019, pp. 1–4. DOI: 10.1109/BIOCAS.2019.8918994 *Cited on pages 16, 41, 56, 60, 92, 95, 107.*

[20] S. Suman, F. A. Hussin, A. S. Malik, *et al.*, "Image enhancement using geometric mean filter and gamma correction for WCE images," in *Neural Information Processing: 21st International Conference, Part III 21*, Springer, 2014, pp. 276–283 *Cited on pages 16, 36, 41, 56, 95.*

[21] Z. Wang, Z. Yao, and Q. Wang, "Improved scheme of estimating motion blur parameters for image restoration," *Digital Signal Processing*, vol. 65, pp. 11–18, 2017 *Cited on page 18.*

[22] E. R. Davies, *Computer and machine vision: theory, algorithms, practicalities*. Academic Press, 2012, pp. 52–54 *Cited on page 18.*

[23] Q. Wang, A. Khanicheh, D. C. Leiner, D. Shafer, and J. Zobel, "Endoscope field of view measurement.," *Biomedical optics express*, vol. 8 3, pp. 1441–1454, 2017 *Cited on page 19.*

[24] L. Mignard-Debise and I. Ihrke, "A vignetting model for light field cameras with an application to light field microscopy," *IEEE Transactions on Computational Imaging*, vol. 5, pp. 585–595, 2019 *Cited on pages 19, 20.*

[25] P. ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," *International telecommunication union*, 1999 *Cited on pages 21, 76.*

[26] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004 *Cited on pages 21, 76, 86.*

[27] M. Zhang, W. Xue, and X. Mou, "Reduced reference image quality assessment based on statistics of edge," in *Digital Photography VII*, vol. 7876, 2011, p. 787 611 *Cited on pages 21, 76.*

[28] A. Mittal, R. Soundararajan, and A. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, pp. 209–212, 2013 *Cited on pages 21, 32, 76, 77, 86.*

[29] A. Mittal, M. Saad, and A. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2016. DOI: 10.1109/TIP.2015.2502725 *Cited on pages 21, 76, 77, 86.*

[30] M. A. Saad, A. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing*, vol. 21, pp. 3339–3352, 2012 *Cited on pages 21, 76, 77.*

[31] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Transactions on communications*, vol. 43, no. 12, pp. 2959–2965, 1995 *Cited on page 22.*

[32] B. Girod, "What's wrong with mean-squared error," *Digital images and human vision*, pp. 207–220, 1993 *Cited on page 22.*

[33] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009 *Cited on page 22.*

[34] C. Lee, S. Cho, J. Choe, T. Jeong, W. Ahn, and E. Lee, "Objective video quality assessment," *Optical engineering*, vol. 45, no. 1, pp. 017 004–017 004, 2006 *Cited on pages 23, 24.*

[35] A. B. Watson, "DCT quantization matrices visually optimized for individual images," in *Human vision, visual processing, and digital display IV*, SPIE, vol. 1913, 1993, pp. 202–216 *Cited on pages 23, 24.*

[36] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE transactions on Information Theory*, vol. 20, no. 4, pp. 525–536, 1974 *Cited on page 24.*

[37] S. J. Daly, "Visible differences predictor: An algorithm for the assessment of image fidelity," in *Human Vision, Visual Processing, and Digital Display III*, SPIE, vol. 1666, 1992, pp. 2–15 *Cited on page 24.*

[38] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proceedings of 1st International Conference on Image Processing*, IEEE, vol. 2, 1994, pp. 982–986 *Cited on page 24.*

[39] J. Lubin, "A visual discrimination model for imaging system design and evaluation," in *Vision Models for Target Detection and Recognition: In Memory of Arthur Menendez*, World Scientific, 1995, pp. 245–283 *Cited on page 24.*

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004 *Cited on page 24.*

[41] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002 *Cited on page 24.*

[42] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2117–2128, 2005 *Cited on page 25.*

[43] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Human vision and electronic imaging X*, SPIE, vol. 5666, 2005, pp. 149–159 *Cited on page 25.*

[44] S. Corchs, F. Gasparini, and R. Schettini, "No reference image quality classification for JPEG-distorted images," *Digital Signal Processing*, vol. 30, pp. 86–100, 2014 *Cited on page 26.*

[45] A. Beghdadi, M.-C. Larabi, A. Bouzerdoum, and K. M. Iftekharuddin, "A survey of perceptual image processing methods," *Signal Processing: Image Communication*, vol. 28, no. 8, pp. 811–831, 2013 *Cited on page 27.*

[46] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, "No-reference image and video quality assessment: A classification and review of recent approaches," *EURASIP Journal on image and Video Processing*, vol. 2014, pp. 1–32, 2014 *Cited on page 27.*

[47] J. Guan, W. Zhang, J. Gu, and H. Ren, "No-reference blur assessment based on edge modeling," *Journal of Visual Communication and Image Representation*, vol. 29, pp. 1–7, 2015 *Cited on page 27.*

[48] Q. Sang, H. Qi, X. Wu, C. Li, and A. C. Bovik, "No-reference image blur index based on singular value curve," *Journal of Visual Communication and Image Representation*, vol. 25, no. 7, pp. 1625–1630, 2014 *Cited on page 27.*

[49] J. Caviedes and S. Gurbuz, "No-reference sharpness metric based on local edge kurtosis," in *Proceedings. International conference on image processing*, IEEE, vol. 3, 2002, pp. III–III *Cited on page 27.*

[50] E. Ong, W. Lin, Z. Lu, *et al.*, "A no-reference quality metric for measuring image blur," in *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.*, Ieee, vol. 1, 2003, pp. 469–472 *Cited on page 27.*

[51] M. Kristan, J. Perš, M. Perše, and S. Kovačič, "A bayes-spectral-entropy-based measure of camera focus using a discrete cosine transform," *Pattern Recognition Letters*, vol. 27, no. 13, pp. 1431–1439, 2006 *Cited on page 27.*

[52] C.-Y. Wee and R. Paramesran, "Image sharpness measure using eigenvalues," in *2008 9th International Conference on Signal Processing*, IEEE, 2008, pp. 840–843 *Cited on page 27.*

[53] S. Varadarajan and L. J. Karam, "An improved perception-based no-reference objective image sharpness metric using iterative edge refinement," in *2008 15th IEEE international conference on image processing*, IEEE, 2008, pp. 401–404 *Cited on page 27.*

[54] N. D. Narvekar and L. J. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *2009 International Workshop on Quality of Multimedia Experience*, IEEE, 2009, pp. 87–91 *Cited on page 28.*

[55] N. G. Sadaka, L. J. Karam, R. Ferzli, and G. P. Abousleman, "A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling," in *2008 15th IEEE International Conference on Image Processing*, IEEE, 2008, pp. 369–372 *Cited on page 28.*

[56] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb)," *IEEE transactions on image processing*, vol. 18, no. 4, pp. 717–728, 2009 *Cited on page 28.*

[57] X. Zhu and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise," in *2009 international workshop on quality of multimedia experience*, IEEE, 2009, pp. 64–69 *Cited on page 28.*

[58] A. Serir, A. Beghdadi, and F. Kerouh, "No-reference blur image quality measure based on multiplicative multiresolution decomposition," *Journal of visual communication and image representation*, vol. 24, no. 7, pp. 911–925, 2013 *Cited on pages 28, 35, 77.*

[59] S. Gabarda and G. Cristóbal, "Blind image quality assessment through anisotropy," *JOSA A*, vol. 24, no. 12, B42–B51, 2007 *Cited on page 28.*

[60] G. Zhai, A. Kaup, J. Wang, and X. Yang, "A dual-model approach to blind quality assessment of noisy images," *APSIPA Transactions on Signal and Information Processing*, vol. 4, e4, 2015 *Cited on page 28.*

[61] K. Rank, M. Lendl, and R. Unbehauen, "Estimation of image noise variance," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 146, no. 2, pp. 80–84, 1999 *Cited on page 28.*

[62] J. Tian and L. Chen, "Image noise estimation using a variation-adaptive evolutionary approach," *IEEE Signal Processing Letters*, vol. 19, no. 7, pp. 395–398, 2012 *Cited on page 28.*

[63] J. Boulanger, C. Kervrann, and P. Bouthemy, "Space-time adaptation for patch-based image sequence restoration," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1096–1102, 2007 *Cited on page 28.*

[64] X. Huang, L. Chen, J. Tian, X. Zhang, and X. Fu, "Blind noisy image quality assessment using block homogeneity," *Computers & Electrical Engineering*, vol. 40, no. 3, pp. 796–807, 2014 *Cited on page 28.*

[65] L. Chen, X. Huang, J. Tian, and X. Fu, "Blind noisy image quality evaluation using a deformable ant colony algorithm," *Optics & Laser Technology*, vol. 57, pp. 265–270, 2014 *Cited on page 28.*

[66] C. Deng, S. Wang, A. C. Bovik, G.-B. Huang, and B. Zhao, "Blind noisy image quality assessment using sub-band kurtosis," *IEEE transactions on cybernetics*, vol. 50, no. 3, pp. 1146–1156, 2019 *Cited on page 28.*

[67] L.-y. Zhou and Z.-b. Zhang, "No-reference image quality assessment based on noise, blurring and blocking effect," *Optik*, vol. 125, no. 19, pp. 5677–5680, 2014 *Cited on page 29.*

[68] L. Liu, B. Liu, H. Huang, and A. C. Bovik, "No-reference image quality assessment based on spatial and spectral entropies," *Signal processing: Image communication*, vol. 29, no. 8, pp. 856–863, 2014 *Cited on page 29.*

[69] X. Li, "Blind image quality assessment," in *International Conference on Image Processing*, IEEE, vol. 1, 2002, pp. I–I *Cited on page 29.*

[70] E. Cohen and Y. Yitzhaky, "No-reference assessment of blur and noise impacts on image quality," *Signal, image and video processing*, vol. 4, pp. 289–302, 2010 *Cited on page 29.*

[71] Z. P. Sazzad, Y. Kawayoke, and Y. Horita, "No reference image quality assessment for JPEG2000 based on spatial features," *Signal Processing: Image Communication*, vol. 23, no. 4, pp. 257–268, 2008 *Cited on page 29.*

[72] Y. Yuan, Q. Guo, and X. Lu, "Image quality assessment: A sparse learning way," *Neurocomputing*, vol. 159, pp. 227–241, 2015 *Cited on page 29.*

[73] A. Mittal, A. Moorthy, and A. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012. DOI: 10.1109/TIP.2012.2214050 *Cited on pages 30, 76, 86.*

[74] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 1, pp. 52–56, 1995 *Cited on page 30.*

[75] L. Liu, H. Dong, H. Huang, and A. C. Bovik, "No-reference image quality assessment in curvelet domain," *Signal Processing: Image Communication*, vol. 29, no. 4, pp. 494–505, 2014 *Cited on page 31.*

[76] M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 583–586, 2010 *Cited on page 31.*

[77] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal processing letters*, vol. 17, no. 5, pp. 513–516, 2010 *Cited on page 31.*

[78] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011 *Cited on page 31.*

[79] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015 *Cited on page 32.*

[80] X. Gao, F. Gao, D. Tao, and X. Li, "Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning," *IEEE Transactions on neural networks and learning systems*, vol. 24, no. 12, 2013 *Cited on page 32.*

[81] C. Li, A. C. Bovik, and X. Wu, "Blind image quality assessment using a general regression neural network," *IEEE Transactions on neural networks*, vol. 22, no. 5, pp. 793–799, 2011 *Cited on pages 32, 33.*

[82] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, pp. 1098–1105 *Cited on page 32.*

[83] L. He, D. Tao, X. Li, and X. Gao, "Sparse representation for blind image quality assessment," in *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2012, pp. 1146–1153 *Cited on page 32.*

[84] Q. Sang, X. Wu, C. Li, and A. C. Bovik, "Blind image quality assessment using a reciprocal singular value curve," *Signal Processing: Image Communication*, vol. 29, no. 10, pp. 1149–1157, 2014 *Cited on page 32.*

[85] H. Tang, N. Joshi, and A. Kapoor, "Blind image quality assessment using semi-supervised rectifier networks," in *proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2877–2884 *Cited on page 32.*

[86] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740 *Cited on page 32.*

[87] R. V. Babu, S. Suresh, and A. Perkis, "No-reference JPEG-image quality assessment using GAP-RBF," *Signal Processing*, vol. 87, no. 6, pp. 1493–1503, 2007 *Cited on page 33.*

[88] D. Ghadiyaram and A. C. Bovik, "Blind image quality assessment on real distorted images using deep belief nets," in *2014 IEEE global conference on signal and information processing (GlobalSIP)*, IEEE, 2014, pp. 946–950 *Cited on page 33.*

[89] Y. Li, L.-M. Po, X. Xu, *et al.*, "No-reference image quality assessment with shearlet transform and deep neural networks," *Neurocomputing*, vol. 154, pp. 94–109, 2015 *Cited on page 33.*

[90] R. N. Forthofer, E. S. Lee, and M. Hernandez, "3 - descriptive methods," in *Biostatistics (Second Edition)*, R. N. Forthofer, E. S. Lee, and M. Hernandez, Eds., Second Edition, San Diego: Academic Press, 2007, pp. 21–69, ISBN: 978-0-12-369492-8. DOI: https://doi.org/10.1016/B978-0-12-369492-8.50008-X *Cited on page 34.*

[91] L. Zhou and Z. Zhang, "No-reference image quality assessment based on noise, blurring and blocking effect," *Optik*, vol. 125, pp. 5677–5680, 2014 *Cited on pages 35, 77.*

[92] Y. Lu, F. Xie, Y. Wu, Z. Jiang, and R. Meng, "No reference uneven illumination assessment for dermoscopy images," *IEEE Signal Processing Letters*, vol. 22, pp. 534–538, 2015 *Cited on pages 35, 77, 81, 86.*

[93] J. Wang, X. Wang, P. Zhang, *et al.*, "Correction of uneven illumination in color microscopic image based on fully convolutional network," *Optics express*, vol. 29, pp. 28 503–28 520, 2021 *Cited on pages 35, 77, 78, 80–82, 86.*

[94] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth International Conference on Computer Vision*, Jan. 1998, pp. 839–846 *Cited on page 36.*

[95] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, Jul. 1990 *Cited on page 36.*

[96] G. Z. Yang, P. Burger, D. N. Firmin, S. R. Underwood, and D. B. Longmore, "Structure adaptive anisotropic filtering," in *Fifth International Conference on Image Processing and its Applications*, Jul. 1995, pp. 717–721 *Cited on page 36.*

[97] S. M. Smith and J. M. Brady, "SUSAN-A new approach to low level image processing," *Int. J. Comput. Vision*, vol. 23, no. 1, pp. 45–78, May 1997 *Cited on page 36.*

[98] S. Haykin and B. Widrow, *Least-mean-square adaptive filters*. Hoboken, N.J.: Wiley-Interscience, 2003 *Cited on page 36.*

[99] L. Shao, H. Zhang, and G. de Haan, "An overview and performance evaluation of classification-based least squares trained filters," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1772–1782, Oct. 2008 *Cited on page 36.*

[100] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 349–366, Feb. 2007 *Cited on page 36.*

[101] X. Li and M. T. Orchard, "New edge-directed interpolation," *IEEE Transactions on Image Processing*, vol. 10, no. 10, pp. 1521–1527, Oct. 2001 *Cited on page 36.*

[102] L. Zhang and X. Wu, "An edge-guided image interpolation algorithm via directional filtering and data fusion," *IEEE Transactions on Image Processing*, vol. 15, no. 8, pp. 2226–2238, Aug. 2006 *Cited on page 36.*

[103] Q. Wang and R. K. Ward, "A new orientation-adaptive interpolation method," *IEEE Transactions on Image Processing*, vol. 16, no. 4, pp. 889–900, Apr. 2007 *Cited on page 36.*

[104] X. Zhang and X. Wu, "Image interpolation by adaptive 2-D autoregressive modeling and soft-decision estimation," *IEEE Transactions on Image Processing*, vol. 17, no. 6, pp. 887–896, Jun. 2008 *Cited on page 36.*

[105] L. G. Shapiro and G. C. Stockman, *Computer Vision*. Pearson, 2001 *Cited on page 36.*

[106] M. Lindenbaum, M. Fischer, and A. Bruckstein, "On gabor's contribution to image enhancement," *Pattern Recognition*, vol. 27, no. 1, pp. 1–8, 1994 *Cited on page 36.*

[107] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, Oct. 1981 *Cited on page 36.*

[108] H. Hou and H. Andrews, "Cubic splines for image interpolation and digital filtering," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 6, pp. 508–517, Oct. 1978 *Cited on page 36.*

[109] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D*, vol. 60, pp. 259–268, Nov. 1992 *Cited on page 37.*

[110] H. Liu, W.-S. Lu, and M. Q. Meng, "Fast algorithms for restoration of color wireless capsule endoscopy images," in *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*, IEEE, 2011, pp. 1–4 *Cited on pages 37, 41, 95.*

[111] O. Scherzer, "Denoising with higher order derivatives of bounded variation and an application to parameter estimation," *Computing*, vol. 60, no. 1, pp. 1–27, Mar. 1998 *Cited on page 38.*

[112] T. Chan, A. Marquina, and P. Mulet, "High-order total variation-based image restoration," *SIAM Journal on Scientific Computing*, vol. 22, no. 2, pp. 503–516, 2000 *Cited on page 38.*

[113] K. Bredies, K. Kunisch, and T. Pock, "Total generalized variation," *SIAM Journal on Imaging Sciences*, vol. 3, no. 3, pp. 492–526, 2010 *Cited on page 38.*

[114] F. Knoll, K. Bredies, T. Pock, and R. Stollberger, "Second order total generalized variation (tgv) for mri," *Magnetic Resonance in Medicine*, vol. 65, no. 2, pp. 480–491, 2011 *Cited on page 38.*

[115] İ. Bayram and M. E. Kamasak, "Directional total variation," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 781–784, Oct. 2012 *Cited on page 38.*

[116] I. Jolliffe, *Principal component analysis*. New York: Springer Verlag, 2002 *Cited on page 40.*

[117] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (gpca)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1945–1959, Oct. 2005 *Cited on page 40.*

[118] S. Bacchelli and S. Papi, "Image denoising using principal component analysis in the wavelet domain," *Journal of Computational and Applied Mathematics*, vol. 189, no. 1, pp. 606–621, 2006 *Cited on page 40.*

[119] Q. Liu, C. Zhang, Q. Guo, H. Xu, and Y. Zhou, "Adaptive sparse coding on PCA dictionary for image denoising," *The Visual Computer*, vol. 32, no. 4, pp. 535–549, Apr. 2016 *Cited on page 40.*

[120] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006 *Cited on page 40.*

[121] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 689–696 *Cited on page 40.*

[122] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010 *Cited on page 40.*

[123] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1553–1564, Mar. 2010 *Cited on page 40.*

[124] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*, 2012, pp. 711–730 *Cited on page 40.*

[125] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, Apr. 2012 *Cited on page 40.*

[126] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 2216–2223 *Cited on page 40.*

[127] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011 *Cited on page 40.*

[128] W. Dong, X. Li, L. Zhang, and G. Shi, "Sparsity-based image denoising via dictionary learning and structural clustering," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2011, pp. 457–464 *Cited on page 40.*

[129] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1620–1630, Apr. 2013 *Cited on page 40.*

[130] V. Jain and S. Seung, "Natural image denoising with convolutional networks," in *Advances in Neural Information Processing Systems 21*, 2009, pp. 769–776 *Cited on page 40.*

[131] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb. 2016 *Cited on page 40.*

[132] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 1646–1654 *Cited on page 40.*

[133] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 1637–1645 *Cited on page 40.*

[134] W. Shi, J. Caballero, F. Huszár, *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 1874–1883 *Cited on page 40.*

[135] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017 *Cited on page 40.*

[136] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2808–2817 *Cited on page 40.*

[137] S. Lefkimmiatis, "Non-local color image denoising with convolutional neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 5882–5891 *Cited on page 40.*

[138] N. Divakar and R. V. Babu, "Image denoising via CNNs: An adversarial approach," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Jul. 2017, pp. 1076–1083 *Cited on page 40.*

[139] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with bm3d?" In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 2392–2399 *Cited on page 41.*

[140] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 341–349 *Cited on page 41.*

[141] F. Agostinelli, M. R. Anderson, and H. Lee, "Adaptive multi-column deep neural networks with application to robust image denoising," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 1493–1501 *Cited on page 41.*

[142] S. W. Zamir, A. Arora, S. Khan, *et al.*, "Cycleisp: Real image restoration via improved data synthesis," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2696–2705, 2020 *Cited on pages 41, 95, 107, 110.*

[143] Z. Yue, Q. Zhao, L. Zhang, and D. Meng, "Dual adversarial network: Toward real-world noise removal and noise generation," *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK*, pp. 41–58, 2020 *Cited on pages 41, 95, 107, 112.*

[144] S. W. Zamir, A. Arora, S. Khan, *et al.*, "Learning enriched features for real image restoration and enhancement," *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 492–511, 2020 *Cited on pages 41, 95, 107–110, 116.*

[145] S. Anwar and N. Barnes, "Real image denoising with feature attention," *Proceedings of the IEEE international conference on computer vision*, pp. 3155–3164, 2019 *Cited on pages 41, 95.*

[146] V. P. Gopi, P. Palanisamy, and S. I. Niwas, "Capsule endoscopic colour image denoising using complex wavelet transform," *Communications in Computer and Information Science*, vol. 292, pp. 220–229, 2012 *Cited on pages 41, 95.*

[147] P. Bojarczak and Z. Lukasik, "Image deblurring–wiener filter versus tsvd approach," *Advances in Electrical and Electronic Engineering*, vol. 6, no. 2, pp. 86–89, 2011 *Cited on page 42.*

[148] M. K. Singh, U. S. Tiwary, and Y.-H. Kim, "An adaptively accelerated lucy-richardson method for image deblurring," *EURASIP Journal on Advances in Signal Processing*, vol. 2008, pp. 1–10, 2007 *Cited on page 42.*

[149] T. F. Chan and C.-K. Wong, "Total variation blind deconvolution," *IEEE transactions on Image Processing*, vol. 7, no. 3, pp. 370–375, 1998 *Cited on pages 42, 95.*

[150] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," in *Acm Siggraph 2006 Papers*, 2006, pp. 787–794 *Cited on page 42.*

[151] K. Zhang, W. Luo, Y. Zhong, *et al.*, "Deblurring by realistic blurring," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2737–2746, 2020 *Cited on pages 43, 95, 108.*

[152] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better," *Proceedings of the IEEE international conference on computer vision*, pp. 8878–8887, 2019 *Cited on pages 43, 93, 95.*

[153] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5978–5986, 2019 *Cited on pages 43, 95, 108.*

[154] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. U. Li, "A general u-shaped transformer for image restoration," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA*, pp. 19–24, 2022 *Cited on pages 43, 96, 107–110, 117.*

[155] H. Liu, W.-S. Lu, and M. Q.-H. Meng, "De-blurring wireless capsule endoscopy images by total variation minimization," in *Proceedings of 2011 IEEE pacific rim conference on communications, computers and signal processing*, IEEE, 2011, pp. 102–106 *Cited on pages 43, 95, 96, 108.*

[156] Y. Wang, C. Cai, and Y. Zou, "Single image super-resolution via adaptive dictionary pair learning for wireless capsule endoscopy image," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, IEEE, 2015, pp. 595–599 *Cited on pages 43, 56, 96.*

[157] L. Peng, S. Liu, D. Xie, S. Zhu, and B. Zeng, "Endoscopic video deblurring via synthesis," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, IEEE, 2017, pp. 1–4 *Cited on pages 43, 96.*

[158] L. Jagannathan and C. Jawahar, "Perspective correction methods for camera based document analysis," in *Proc. First Int. Workshop on Camera-based Document Analysis and Recognition*, International Institute of Information Technology Hyderabad, India, 2005, pp. 148–154 *Cited on page 44.*

[159] D. Pavić, V. Schönefeld, and L. Kobbelt, "Interactive image completion with perspective correction," *The Visual Computer*, vol. 22, pp. 671–681, 2006 *Cited on page 44.*

[160] N. Dey, "Uneven illumination correction of digital images: A survey of the state-of-the-art," *Optik*, vol. 183, pp. 483–495, 2019 *Cited on page 45.*

[161] R. Baumann, C. Blackwell, and W. B. Seales, "Automatic perspective correction of manuscript images," in *The Outreach of Digital Libraries: A Globalized Resource Network: 14th International Conference on Asia-Pacific Digital Libraries, ICADL 2012, Taipei, Taiwan, November 12-15, 2012, Proceedings 14*, Springer, 2012, pp. 11–18 *Cited on page 45.*

[162] S. Lu, B. M. Chen, and C. C. Ko, "Perspective rectification of document images using fuzzy set and morphological operations," *Image and Vision Computing*, vol. 23, no. 5, pp. 541–553, 2005 *Cited on page 45.*

[163] W. Zhang, X. Li, and X. Ma, "Perspective correction method for chinese document images," in *2008 International Symposium on Intelligent Information Technology Application Workshops*, 2008, pp. 467–470. DOI: 10.1109/IITA.Workshops.2008.148 *Cited on page 45.*

[164] M. Golpardaz and H. Nezamabadi-Pour, "Perspective rectification and skew correction in camera-based farsi document images," in *2011 7th Iranian Conference on Machine Vision and Image Processing*, 2011, pp. 1–5. DOI: 10.1109/IranianMVIP.2011.6121613 *Cited on page 45.*

[165] F. W. Leong, M. Brady, and J. O. McGee, "Correction of uneven illumination (vignetting) in digital microscopy images," *Journal of clinical pathology*, vol. 56, no. 8, pp. 619–621, 2003 *Cited on pages 45, 96.*

[166] J. Wang, X. Wang, P. Zhang, *et al.*, "Correction of uneven illumination in color microscopic image based on fully convolutional network," *Optics express*, vol. 29, no. 18, pp. 28 503–28 520, 2021 *Cited on pages 45, 96, 109.*

[167] R. Han, C. Tang, M. Xu, and Z. Lei, "A retinex-based variational model for noise suppression and non-uniform illumination correction in corneal confocal microscopy images," *Physics in Medicine and Biology*, 2022 *Cited on pages 45, 96.*

[168] X. Cao, S. Rong, Y. Liu, T. Li, Q. Wang, and B. He, "NUICNet: Non-uniform illumination correction for underwater image using fully convolutional network," *IEEE Access*, vol. 8, pp. 109 989–110 002, 2020 *Cited on page 46.*

[169] S. S. Sankpal, S. S. Deshpande, *et al.*, "Nonuniform illumination correction algorithm for underwater images using maximum likelihood estimation method," *Journal of Engineering*, vol. 2016, 2016 *Cited on page 46.*

[170] F. Piccinini and A. Bevilacqua, "Colour vignetting correction for microscopy image mosaics used for quantitative analyses," *BioMed research international*, vol. 2018, 2018 *Cited on page 46.*

[171] M. Long, Z. Li, X. Xie, G. Li, and Z. Wang, "Adaptive image enhancement based on guide image and fraction-power transformation for wireless capsule endoscopy," *IEEE transactions on biomedical circuits and systems*, vol. 12, no. 5, pp. 993–1003, 2018 *Cited on pages 46, 96, 109.*

[172] K. Zhan, J. Teng, J. Shi, Q. Li, and M. Wang, "Feature-linking model for image enhancement," *Neural computation*, vol. 28, no. 6, pp. 1072–1100, 2016 *Cited on pages 46, 96.*

[173] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Transactions on image processing*, vol. 26, no. 2, pp. 982–993, 2016 *Cited on pages 47, 96, 109, 110.*

[174] Y. Jiang, X. Gong, D. Liu, *et al.*, "Enlightengan: Deep light enhancement without paired supervision," *IEEE transactions on image processing*, vol. 30, pp. 2340–2349, 2021 *Cited on pages 47, 93, 96, 109, 110.*

[175] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," *arXiv preprint arXiv:1808.04560*, 2018 *Cited on pages 47, 96, 109.*

[176] O. H. Maghsoudi and H. Soltanian-Zadeh, "Detection of abnormalities in wireless capsule endoscopy frames using local fuzzy patterns," in *2013 20th Iranian Conference on Biomedical Engineering (ICBME)*, IEEE, 2013, pp. 286–291 *Cited on page 48.*

[177] D. K. Iakovidis, S. V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, and V. P. Plagianakos, "Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification," *IEEE transactions on medical imaging*, vol. 37, no. 10, pp. 2196–2210, 2018 *Cited on page 48.*

[178] X. Guo, L. Zhang, Y. Hao, L. Zhang, Z. Liu, and J. Liu, "Multiple abnormality classification in wireless capsule endoscopy images based on efficientnet using attention mechanism," *Review of Scientific Instruments*, vol. 92, no. 9, 2021 *Cited on page 48.*

[179] F. Noya, M. A. Álvarez-González, and R. Benitez, "Automated angiodysplasia detection from wireless capsule endoscopy," in *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, IEEE, 2017, pp. 3158–3161 *Cited on page 48.*

[180] R. Leenhardt, P. Vasseur, C. Li, *et al.*, "A neural network algorithm for detection of gi angiectasia during small-bowel capsule endoscopy," *Gastrointestinal endoscopy*, vol. 89, no. 1, pp. 189–194, 2019 *Cited on page 48.*

[181] M. A. Usman, G. B. Satrya, M. R. Usman, and S. Y. Shin, "Detection of small colon bleeding in wireless capsule endoscopy videos," *Computerized Medical Imaging and Graphics*, vol. 54, pp. 16–26, 2016 *Cited on page 49.*

[182] G. Pan, F. Xu, and J. Chen, "A novel algorithm for color similarity measurement and the application for bleeding detection in WCE," *IJ Image, Graphics and Signal Processing*, vol. 5, pp. 1–7, 2011 *Cited on page 49.*

[183] S. Sainju, F. M. Bui, and K. A. Wahid, "Automated bleeding detection in capsule endoscopy videos using statistical features and region growing," *Journal of medical systems*, vol. 38, pp. 1–11, 2014 *Cited on page 49.*

[184] T. Ghosh, S. A. Fattah, and K. A. Wahid, "Chobs: Color histogram of block statistics for automatic bleeding detection in wireless capsule endoscopy video," *IEEE journal of translational engineering in health and medicine*, vol. 6, pp. 1–12, 2018 *Cited on page 49.*

[185] I. N. Figueiredo, S. Kumar, C. Leal, and P. N. Figueiredo, "Computer-assisted bleeding detection in wireless capsule endoscopy images," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 1, no. 4, pp. 198–210, 2013 *Cited on page 49.*

[186] A. K. Kundu, S. A. Fattah, M. N. Rizve, *et al.*, "An automatic bleeding frame and region detection scheme for wireless capsule endoscopy videos based on interplane intensity variation profile in normalized rgb color space," *Journal of healthcare engineering*, vol. 2018, 2018 *Cited on page 49.*

[187] T. Ghosh, S. A. Fattah, and K. A. Wahid, "Automatic computer aided bleeding detection scheme for wireless capsule endoscopy (WCE) video based on higher and lower order statistical features in a composite color," *Journal of Medical and Biological Engineering*, vol. 38, pp. 482–496, 2018 *Cited on page 49.*

[188] Y. Fu, W. Zhang, M. Mandal, and M. Q.-H. Meng, "Computer-aided bleeding detection in WCE video," *IEEE journal of biomedical and health informatics*, vol. 18, no. 2, pp. 636–642, 2013 *Cited on page 49.*

[189] X. Jia, L. Cai, J. Liu, W. Dai, and M. Q.-H. Meng, "Gi bleeding detection in wireless capsule endoscopy images based on pattern recognition and a mapreduce framework," in *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, IEEE, 2016, pp. 266–271 *Cited on page 49.*

[190] O. Bchir, M. M. Ben Ismail, and N. AlZahrani, "Multiple bleeding detection in wireless capsule endoscopy," *Signal, Image and Video Processing*, vol. 13, pp. 121–126, 2019 *Cited on page 49.*

[191] X. Xing, Y. Yuan, X. Jia, and M. Q.-H. Meng, "A saliency-aware hybrid dense network for bleeding detection in wireless capsule endoscopy images," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, 2019, pp. 104–107 *Cited on page 49.*

[192] M. Hajabdollahi, R. Esfandiarpoor, P. Khadivi, *et al.*, "Segmentation of bleeding regions in wireless capsule endoscopy for detection of informative frames," *Biomedical Signal Processing and Control*, vol. 53, p. 101 565, 2019 *Cited on page 49.*

[193] S. Hwang and M. E. Celebi, "Polyp detection in wireless capsule endoscopy videos based on image segmentation and geometric feature," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 678–681 *Cited on page 49.*

[194] A. Karargyris and N. Bourbakis, "Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos," *IEEE Transactions on biomedical engineering*, vol. 58, no. 10, pp. 2777–2786, 2011 *Cited on page 49.*

[195] Y. Yuan and M. Q.-H. Meng, "A novel feature for polyp detection in wireless capsule endoscopy images," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2014, pp. 5010–5015 *Cited on page 49.*

[196] L. Gueye, S. Yildirim-Yayilgan, F. A. Cheikh, and I. Balasingham, "Automatic detection of colonoscopic anomalies using capsule endoscopy," in *2015 IEEE international conference on image processing (ICIP)*, IEEE, 2015, pp. 1061–1064 *Cited on page 49.*

[197] Y. Iwahori, A. Hattori, Y. Adachi, M. K. Bhuyan, R. J. Woodham, and K. Kasugai, "Automatic detection of polyp using hessian filter and HOG features," *Procedia computer science*, vol. 60, pp. 730–739, 2015 *Cited on page 49.*

[198] M. El Ansari and S. Charfi, "Computer-aided system for polyp detection in wireless capsule endoscopy images," in *2017 International Conference on Wireless Networks and Mobile Communications (WINCOM)*, IEEE, 2017, pp. 1–6 *Cited on page 49.*

[199] S.-H. Bae and K.-J. Yoon, "Polyp detection via imbalanced learning and discriminative feature learning," *IEEE transactions on medical imaging*, vol. 34, no. 11, pp. 2379–2393, 2015 *Cited on page 49.*

[200] Y. Yuan, B. Li, and M. Q.-H. Meng, "Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images," *IEEE Transactions on automation science and engineering*, vol. 13, no. 2, pp. 529–535, 2015 *Cited on page 49.*

[201] M. Billah, S. Waheed, M. M. Rahman, *et al.*, "An automatic gastrointestinal polyp detection system in video endoscopy using fusion of color wavelet and convolutional neural network features," *International journal of biomedical imaging*, vol. 2017, 2017 *Cited on page 49.*

[202] E. S. Nadimi, M. M. Buijs, J. Herp, *et al.*, "Application of deep learning for autonomous detection and localization of colorectal polyps in wireless colon capsule endoscopy," *Computers & Electrical Engineering*, vol. 81, p. 106 531, 2020 *Cited on page 49.*

[203] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection using global geometric constraints and local intensity variation patterns," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II 17*, Springer, 2014, pp. 179–187 *Cited on page 49.*

[204] E. Ribeiro, A. Uhl, G. Wimmer, M. Häfner, *et al.*, "Exploring deep learning and transfer learning for colonic polyp classification," *Computational and mathematical methods in medicine*, 2016 *Cited on page 49.*

[205] R. Zhang, Y. Zheng, C. C. Poon, D. Shen, and J. Y. Lau, "Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker," *Pattern recognition*, vol. 83, pp. 209–219, 2018 *Cited on page 49.*

[206] S. Sornapudi, F. Meng, and S. Yi, "Region-based automated localization of colonoscopy and wireless capsule endoscopy polyps," *Applied Sciences*, vol. 9, no. 12, p. 2404, 2019 *Cited on page 49.*

[207] Y. Yuan and M. Q.-H. Meng, "Deep learning for polyp recognition in wireless capsule endoscopy images," *Medical physics*, vol. 44, no. 4, pp. 1379–1389, 2017 *Cited on page 50.*

[208] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE transactions on information technology in biomedicine*, vol. 7, no. 3, pp. 141–152, 2003 *Cited on page 50.*

[209] B. Li and M. Q.-H. Meng, "Tumor recognition in wireless capsule endoscopy images using textural features and svm-based feature selection," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, no. 3, pp. 323–329, 2012 *Cited on page 50.*

[210] V. Faghih Dinevari, G. Karimian Khosroshahi, M. Zolfy Lighvan, *et al.*, "Singular value decomposition based features for automatic tumor detection in wireless capsule endoscopy images," *Applied bionics and biomechanics*, vol. 2016, 2016 *Cited on page 50.*

[211] G. Liu, G. Yan, S. Kuang, and Y. Wang, "Detection of small bowel tumor based on multi-scale curvelet analysis and fractal technology in capsule endoscopy," *Computers in biology and medicine*, vol. 70, pp. 131–138, 2016 *Cited on page 50.*

[212] M. M. Martins, D. J. Barbosa, J. Ramos, and C. S. Lima, "Small bowel tumors detection in capsule endoscopy by gaussian modeling of color curvelet covariance coefficients," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, IEEE, 2010, pp. 5557–5560 *Cited on page 50.*

[213] P. M. Vieira, J. Ramos, and C. S. Lima, "Automatic detection of small bowel tumors in endoscopic capsule images by roi selection based on discarded lightness information," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2015, pp. 3025–3028 *Cited on page 50.*

[214] M. Alizadeh, O. H. Maghsoudi, K. Sharzehi, H. R. Hemati, A. K. Asl, and A. Talebpour, "Detection of small bowel tumor in wireless capsule endoscopy images using an adaptive neuro-fuzzy inference system," *Journal of biomedical research*, vol. 31, no. 5, p. 419, 2017 *Cited on page 50.*

[215] A. Karargyris and N. Bourbakis, "Identification of ulcers in wireless capsule endoscopy videos," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, IEEE, 2009, pp. 554–557 *Cited on page 50.*

[216] A. Eid, V. S. Charisis, L. J. Hadjileontiadis, and G. D. Sergiadis, "A curvelet-based lacunarity approach for ulcer detection from wireless capsule endoscopy images," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, IEEE, 2013, pp. 273–278 *Cited on page 50.*

[217] M. Souaidi, A. A. Abdelouahed, and M. El Ansari, "Multi-scale completed local binary patterns for ulcer detection in wireless capsule endoscopy images,"

Multimedia Tools and Applications, vol. 78, pp. 13 091–13 108, 2019 *Cited on page 50.*

[218] S. Wang, Y. Xing, L. Zhang, H. Gao, and H. Zhang, "A systematic evaluation and optimization of automatic detection of ulcers in wireless capsule endoscopy on a large dataset using deep convolutional neural networks," *Physics in Medicine & Biology*, vol. 64, no. 23, p. 235 014, 2019 *Cited on page 50.*

[219] E. Klang, Y. Barash, R. Y. Margalit, *et al.*, "Deep learning algorithms for automated detection of crohn's disease ulcers by video capsule endoscopy," *Gastrointestinal endoscopy*, vol. 91, no. 3, pp. 606–613, 2020 *Cited on page 50.*

[220] S. Charfi, M. El Ansari, and I. Balasingham, "Computer-aided diagnosis system for ulcer detection in wireless capsule endoscopy images," *IET Image Processing*, vol. 13, no. 6, pp. 1023–1030, 2019 *Cited on page 50.*

[221] A. K. Sekuboyina, S. T. Devarakonda, and C. S. Seelamantula, "A convolutional neural network approach for abnormality detection in wireless capsule endoscopy," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, IEEE, 2017, pp. 1057–1060 *Cited on page 51.*

[222] V. S. Sadasivan and C. S. Seelamantula, "High accuracy patch-level classification of wireless capsule endoscopy images using a convolutional neural network," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE, 2019, pp. 96–99 *Cited on pages 51, 126, 150, 151.*

[223] Y. Yuan, B. Li, and M. Q.-H. Meng, "WCE abnormality detection based on saliency and adaptive locality-constrained linear coding," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 1, pp. 149–159, 2016 *Cited on page 51.*

[224] R. Nawarathna, J. Oh, J. Muthukudage, *et al.*, "Abnormal image detection in endoscopy videos using a filter bank and local binary patterns," *Neurocomputing*, vol. 144, pp. 70–91, 2014 *Cited on page 51.*

[225] S. Nadeem, M. A. Tahir, S. S. A. Naqvi, and M. Zaid, "Ensemble of texture and deep learning features for finding abnormalities in the gastro-intestinal tract," in *Computational Collective Intelligence: 10th International Conference, ICCCI 2018, Bristol, UK, September 5-7, 2018, Proceedings, Part II 10*, Springer, 2018, pp. 469–478 *Cited on pages 51, 125.*

[226] L. Lan, C. Ye, C. Wang, and S. Zhou, "Deep convolutional neural networks for WCE abnormality detection: CNN architecture, region proposal and transfer learning," *IEEE Access*, vol. 7, pp. 30 017–30 032, 2019 *Cited on pages 51, 125, 150, 151.*

[227] X. Xing, Y. Yuan, and M. Q.-H. Meng, "Zoom in lesions for better diagnosis: Attention guided deformation network for WCE image classification," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4047–4059, 2020 *Cited on pages 52, 126, 150, 151.*

[228] X. Guo and Y. Yuan, "Triple ANet: Adaptive abnormal-aware attention network for WCE image classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, Springer, 2019, pp. 293–301 *Cited on page 52.*

[229] Q. Zhao, W. Yang, and Q. Liao, "Adasan: Adaptive cosine similarity self-attention network for gastrointestinal endoscopy image classification," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2021, pp. 1855–1859 *Cited on page 52.*

[230] M. W. Alam, M. H. A. Sohag, A. H. Khan, T. Sultana, and K. A. Wahid, "Iot-based intelligent capsule endoscopy system: A technical review," *Intelligent Data Analysis for Biomedical Applications*, pp. 1–20, 2019 *Cited on pages 56, 92.*

[231] H. Liu, W.-S. Lu, and M. Q.-H. Meng, "De-blurring wireless capsule endoscopy images by total variation minimization," in *Proceedings of 2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 2011, pp. 102–106. DOI: 10.1109/PACRIM.2011.6032875 *Cited on pages 56, 92.*

[232] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012 *Cited on page 56.*

[233] V. Hosu, F. Hahn, M. Jenadeleh, *et al.*, "The konstanz natural video database (KoNViD-1k)," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–6 *Cited on page 56.*

[234] N. Ponomarenko, O. Ieremeiev, V. Lukin, *et al.*, "Color image database TID2013: Peculiarities and preliminary results," in *European workshop on visual information processing (EUVIP)*, IEEE, 2013, pp. 106–111 *Cited on page 56.*

[235] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of electronic imaging*, vol. 19, no. 1, pp. 011 006–011 006, 2010 *Cited on page 56.*

[236] X. Liu, M. Pedersen, and J. Y. Hardeberg, "CID: IQ–a new image quality database," in *Image and Signal Processing: 6th International Conference, Cherbourg, France*, Springer, 2014, pp. 193–202 *Cited on page 56.*

[237] H. Fu, B. Wang, J. Shen, *et al.*, "Evaluation of retinal image quality assessment networks in different color-spaces," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, Springer, 2019, pp. 48–56 *Cited on page 56.*

[238] Z. A. Khan, A. Beghdadi, F. A. Cheikh, *et al.*, "Towards a video quality assessment based framework for enhancement of laparoscopic videos," in *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment*, vol. 11316, 2020, pp. 129–136 *Cited on page 56.*

[239] J. Immerkaer, "Fast noise variance estimation," *Computer vision and image understanding*, vol. 64, no. 2, pp. 300–302, 1996 *Cited on page 59.*

[240] D. D. Muresan and T. W. Parks, "Adaptive principal components and image denoising," *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, vol. 1, pp. I–101, 2003 *Cited on page 59.*

[241] A. Chetouani, A. Beghdadi, and M. Deriche, "A new reference-free image quality index for blur estimation in the frequency domain," in *2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, IEEE, 2009, pp. 155–159 *Cited on pages 59, 82.*

[242] T.-S. Nguyen, J. Chaussard, M. Luong, H. Zaag, and A. Beghdadi, "A no-reference measure for uneven illumination assessment on laparoscopic images," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2022, pp. 4103–4107 *Cited on page 60.*

[243] J. Park, Y. Hwang, J.-H. Yoon, *et al.*, "Recent development of computer vision technology to improve capsule endoscopy," *Clinical endoscopy*, vol. 52, pp. 328–333, 2019 *Cited on page 62.*

[244] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI'81: 7th international joint conference on Artificial intelligence*, vol. 2, 1981, pp. 674–679 *Cited on page 62.*

[245] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979 *Cited on page 62.*

[246] S. J. Fletcher and M. Zupanski, "A hybrid multivariate normal and lognormal distribution for data assimilation," *Atmospheric Science Letters*, vol. 7, no. 2, pp. 43–46, 2006 *Cited on page 64.*

[247] J. Clark, "The ishihara test for color blindness.," in *American Journal of Physiological Optics*, vol. 5, pp. 269–276 *Cited on page 69.*

[248] ITU-T, "Subjective video quality assessment methods for multimedia applications," *Recommendation P910, International telecommunication union*, 2008 *Cited on page 69.*

[249] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014 *Cited on page 72.*

[250] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255 *Cited on page 72.*

[251] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *Journal of machine learning research*, vol. 9, no. 11, 2008 *Cited on page 72.*

[252] J. A. Forrester, N. Boyd, J. Fitzgerald, I. Wilson, A. Bekele, and T. Weiser, "Impact of surgical lighting on intraoperative safety in low-resource settings: A cross-sectional survey of surgical providers," *World Journal of Surgery*, vol. 41, pp. 3055–3065, 2017 *Cited on page 76.*

[253] H. Hemphälä., W. Osterhaus, P. O. Larsson, J. Borell, and P. Nylén, "Towards better lighting recommendations for open surgery," in *Lighting Research & Technology*, 2020, pp. 856–882 *Cited on page 76.*

[254] L. Hussain and A. Alamry, "Correction of non-uniform illumination for biological images using morphological operation assessing with statistical features quality," *Ibn Al-Haitham Journal For Pure And Applied Science*, vol. 29, pp. 81–90, 2017 *Cited on page 76.*

[255] R. Naseem, Z. Khan, N. Satpute, A. Beghdadi, F. Cheikh, and J. Olivares, "Cross-modality guided contrast enhancement for improved liver tumor image segmentation," *IEEE Access*, vol. 9, pp. 118 154–118 167, 2021 *Cited on page 76.*

[256] R. Palomar, F. Cheikh, B. Edwin, A. Beghdadi, and O. Elle, "Surface reconstruction for planning and navigation of liver resections," *Computerized medical imaging and graphics*, vol. 53, pp. 30–42, 2016 *Cited on page 76.*

[257] Z. A. Khan, A. Beghdadi, F. Cheikh, *et al.*, "Towards a video quality assessment based framework for enhancement of laparoscopic videos," in *Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment, Houston, TX, USA*, ser. SPIE Proceedings, vol. 11316, February 15-20, 2020, pp. 129–136 *Cited on pages 78, 81, 86.*

[258] F. J. W.-M. Leong, M. Brady, and J. McGee, "Correction of uneven illumination (vignetting) in digital microscopy images," *Journal of Clinical Pathology*, vol. 56, pp. 619–621, 2003 *Cited on page 82.*

[259] J.-F. Lalonde, A. Efros, and S. Narasimhan, "Estimating natural illumination from a single outdoor image," *12th IEEE International Conference on Computer Vision*, pp. 183–190, 2009 *Cited on page 83.*

[260] T. Kim and K. Hong, "A practical single image based approach for estimating illumination distribution from shadows," *10th IEEE International Conference on Computer Vision (ICCV)*, vol. 1, pp. 266–271, 2005 *Cited on page 83.*

[261] H. Li, L. Zhang, and H. Shen, "A perceptually inspired variational method for the uneven intensity correction of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 8, pp. 3053–3065, 2012. DOI: 10.1109/TGRS.2011.2178075 *Cited on page 83.*

[262] S.-C. Pei, Y. Hsiao, M. Tzeng, and F. Chang, "Uneven illumination removal and image enhancement using empirical mode decomposition," *Journal of Electronic Imaging*, vol. 22, pp. 43 037–43 037, 2013 *Cited on page 83.*

[263] M. Outtas, L. Zhang, O. Déforges, W. Hamidouche, A. Serir, and C. Cavaro-Ménard, "A study on the usability of opinion-unaware no-reference natural image quality metrics in the context of medical images," *2016 International Symposium on Signal, Image, Video and Communications (ISIVC)*, pp. 308–313, 2016 *Cited on page 86.*

[264] M. Halpern and J. Harold, *Atlas of capsule endoscopy*. Given Imaging, 2002 *Cited on page 92.*

[265] P. Swain and A. Fritscher-Ravens, "Role of video endoscopy in managing small bowel disease," *Gut*, vol. 53, no. 12, pp. 1866–1875, 2004 *Cited on page 92.*

[266] V. Vani and K. M. Prashanth, "Image enhancement of wireless capsule endoscopy frames using image fusion technique," *IETE Journal of Research*, vol. 67, no. 4, pp. 463–475, 2021 *Cited on page 92.*

[267] M. Long, Z. Li, X. Xie, G. Li, and Z. Wang, "Adaptive image enhancement based on guide image and fraction-power transformation for wireless capsule endoscopy," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 5, pp. 993–1003, 2018. DOI: 10.1109/TBCAS.2018.2869530 *Cited on page 92.*

[268] S. van Vliet, A. Sobiecki, and A. C. Telea, "Joint brightness and tone stabilization of capsule endoscopy videos.," in *VISIGRAPP (4: VISAPP)*, 2018, pp. 101–112 *Cited on page 92.*

[269] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," *Proceedings of the 27th ACM International Conference on Multimedia*, 2019 *Cited on page 93.*

[270] H. Bhojwani, V. Bhavsar, R. Gajjar, and M. I. Patel, "Image resolution enhancement using convolutional autoencoders with skip connections," *International Conference on Range Technology (ICORT)*, pp. 1–5, 2021 *Cited on page 93.*

[271] Y. Cai and K. T. U, "Low-light image enhancement based on modified u-net," *2019 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*, pp. 1–7, 2019 *Cited on page 93.*

[272] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017 *Cited on page 93.*

[273] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14, Montreal, Canada: MIT Press, 2014, pp. 3104–3112 *Cited on page 93.*

[274] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196 *Cited on page 94.*

[275] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007 *Cited on pages 95, 107.*

[276] K. Bredies, K. Kunisch, and T. Pock, "Total generalized variation," *SIAM Journal on Imaging Sciences*, vol. 3, no. 3, pp. 492–526, 2010 *Cited on page 95.*

[277] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006 *Cited on page 95.*

[278] Y. Li, A. W. Yu, T. Meng, *et al.*, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 182–17 191 *Cited on pages 98, 99.*

[279] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," *Advances in neural information processing systems*, vol. 16, 2003 *Cited on page 99.*

[280] O. Konur, D. Kingma, and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.(ICLR)*, 2015, pp. 1–15 *Cited on page 102.*

[281] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979 *Cited on page 103.*

[282] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987 *Cited on page 103.*

[283] Z. Yue, H. Yong, Q. Zhao, D. Meng, and L. Zhang, "Variational denoising network: Toward blind noise modeling and removal," *Advances in neural information processing systems*, vol. 32, 2019 *Cited on page 107.*

[284] A. Mehri, P. B. Ardakani, and A. D. Sappa, "MPRNet: Multi-path residual network for lightweight image super resolution," *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 2704–2713, 2021 *Cited on pages 107, 108.*

[285] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better," *IEEE International Conference on Computer Vision (ICCV)*, pp. 8877–8886, 2019 *Cited on page 108.*

[286] H. A. David, *The method of paired comparisons*. London, 1963, vol. 12 *Cited on page 117.*

[287] R. L. Siegel, K. D. Miller, A. Goding Sauer, *et al.*, "Colorectal cancer statistics, 2020," *CA: a cancer journal for clinicians*, vol. 70, no. 3, pp. 145–164, 2020 *Cited on page 122.*

[288] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, no. 6785, pp. 417–417, 2000 *Cited on page 122.*

[289] A. Karargyris and N. Bourbakis, "Wireless capsule endoscopy and endoscopic imaging: A survey on various methodologies presented," *IEEE Engineering in medicine and biology magazine*, vol. 29, no. 1, pp. 72–83, 2010 *Cited on page 122.*

[290] M. J. Willemink, W. A. Koszek, C. Hardell, *et al.*, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020 *Cited on page 122.*

[291] M. F. A. Hady and F. Schwenker, "Semi-supervised learning," *Handbook on Neural Information Processing*, pp. 215–239, 2013 *Cited on page 122.*

[292] A. K. Sekuboyina, S. T. Devarakonda, and C. S. Seelamantula, "A convolutional neural network approach for abnormality detection in wireless capsule endoscopy," *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 1057–1060, 2017 *Cited on pages 125, 150, 151.*

[293] C. Sindhu and V. Valsan, "Automatic detection of colonic polyps and tumor in wireless capsule endoscopy images using hybrid patch extraction and supervised classification," *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pp. 1–5, 2017 *Cited on pages 125, 150, 151.*

[294] M. Sharif, M. Attique Khan, M. Rashid, M. Yasmin, F. Afza, and U. J. Tanik, "Deep CNN and geometric features-based gastrointestinal tract diseases detection and classification from wireless capsule endoscopy images," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 33, no. 4, pp. 577–599, 2021 *Cited on pages 125, 150, 151.*

[295] Y. Cao, W. Yang, K. Chen, Y. Ren, and Q. Liao, "Capsule endoscopy image classification with deep convolutional neural networks," *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pp. 1584–1588, 2018 *Cited on pages 125, 150, 151.*

[296] X. Xing, Y. Yuan, and M. Q.-H. Meng, "Diagnose like a clinician: Third-order attention guided lesion amplification network for WCE image classification," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, pp. 10 145–10 151 *Cited on pages 126, 150, 151.*

[297] X. Guo and Y. Yuan, "Triple ANet: Adaptive abnormal-aware attention network for WCE image classification," *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pp. 293–301, 2019 *Cited on pages 126, 150, 151.*

[298] X. Guo and Y. Yuan, "Semi-supervised WCE image classification with adaptive aggregated attention," *Medical Image Analysis*, vol. 64, p. 101 733, 2020 *Cited on pages 126, 150, 151.*

[299] Q. Zhao, W. Yang, and Q. Liao, "Adasan: Adaptive cosine similarity self-attention network for gastrointestinal endoscopy image classification," *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1855–1859, 2021 *Cited on pages 126, 150, 151.*

[300] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *International conference on machine learning*, pp. 1597–1607, 2020 *Cited on pages 128, 129.*

[301] J.-B. Grill, F. Strub, F. Altché, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020 *Cited on pages 129, 132.*

[302] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020 *Cited on pages 130, 133.*

[303] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018 *Cited on pages 131, 134.*

[304] Z. Xie, Y. Lin, Z. Yao, *et al.*, "Self-supervised learning with swin transformers," *arXiv preprint arXiv:2105.04553*, 2021 *Cited on page 132.*

[305] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020 *Cited on page 132.*

[306] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10 012–10 022, 2021 *Cited on pages 136–138.*

[307] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers and distillation through attention (2020). doi: 10.48550," *arxiv*, 2012 *Cited on page 137.*

[308] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020 *Cited on page 138.*

[309] Z. Liu, H. Hu, Y. Lin, *et al.*, "Swin transformer v2: Scaling up capacity and resolution," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12 009–12 019, 2022 *Cited on pages 139, 140.*