



HAL
open science

Automatic quantification of ocular dryness by artificial intelligence in the context of Sjögren's syndrome

Ikram Brahim

► **To cite this version:**

Ikram Brahim. Automatic quantification of ocular dryness by artificial intelligence in the context of Sjögren's syndrome. Human health and pathology. Université de Bretagne occidentale - Brest, 2022. English. NNT : 2022BRES0112 . tel-04095848

HAL Id: tel-04095848

<https://theses.hal.science/tel-04095848>

Submitted on 12 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ABSTRACT

Sjögren’s syndrome is an immune system disorder with two common symptoms, dry eyes and a dry mouth. The discomfort of dry eye symptoms affects daily lives, results in 30% activity impairment and affects 95% of Sjögren patients [1]. Dry eye disease (DED) is also an independent multifactorial disorder with a prevalence of up to 50% [2]. The ocular surface inflammation causes discomfort, fatigue and overall, a lower quality of life [2, 3]. Traditional therapies help manage the symptoms and avoid permanent damage. Hence, it is pivotal to grade and follow the development of DED. A common drawback in existing methods that diagnose and quantify DED is reproducibility, invasivity and inaccuracy. We reviewed classical methods and those that incorporate automation to measure the extent of DED [4]. The study showed that DED has yet to benefit from what Artificial Intelligence (AI) has to offer. Using slit-lamp examinations of the ocular surface we aimed to improve the quantification of the Oxford score [5]. Our proposed method uses unsupervised learning to register frames from the examinations to a common coordinate system. By learning the camera motion and depth simultaneously we are able to track the ocular surface in 3-D, compensate for eye motion and visualise the full eye. The light source attached to the camera is a challenge and a disturbance when learning egomotion. This was solved through semantic segmentation and adding a new supervision signal: semantic reconstruction loss. We also used the advantage of estimating the shape of the eye as prior knowledge we could include as a constraint. This was implemented through a shape fitting loss; the shapes being two spheres intersecting each other. Our registration showed quantitative and qualitative improvement with each contribution. We also calculated the inter-rater reliability of the punctate dots (damaged areas) annotations. Our method came closest to what can be considered human error. The proposed registration method was also used for a pre-processing task, frame selection. Once applied to automated Oxford score classification, our method improved the results as well. The improvement validates that the strong color/illumination variances present in the examinations are a disturbance for any deep learning task. We overcame this via our contributions and proposed method.

RÉSUMÉ

Le syndrome de Sjögren est une maladie du système immunitaire dont les deux symptômes communs sont la sécheresse des yeux et celle de la bouche. La gêne occasionnée par les symptômes de sécheresse oculaire affecte la vie quotidienne, entraîne une diminution de 30% des activités et touche 95% des patients atteints du syndrome de Sjögren [1]. Le syndrome de l'œil sec (SOS) est également un trouble multifactoriel indépendant dont la prévalence peut atteindre 50% [2]. L'inflammation de la surface oculaire entraîne une gêne, une fatigue et, globalement, une baisse de la qualité de vie [2, 3]. Les thérapies traditionnelles permettent de gérer les symptômes et d'éviter les dommages permanents. Il est donc essentiel de classer et de suivre l'évolution du SOS. Les méthodes existantes qui permettent de diagnostiquer et de quantifier le SOS présentent des inconvénients communs : la reproductibilité, l'invasivité et l'imprécision. Nous avons passé en revue les méthodes classiques et celles qui intègrent l'automatisation pour mesurer l'étendue du SOS : [4]. Cette étude a montré que le SOS n'a pas encore bénéficié de ce que l'intelligence artificielle (IA) a à offrir. En utilisant des examens de la surface oculaire à la lampe à fente, nous avons cherché à améliorer la quantification du score d'Oxford [5]. La méthode que nous proposons utilise l'apprentissage non supervisé pour recalibrer les images des examens dans un système de coordonnées commun. En apprenant simultanément le mouvement de la caméra et la profondeur, nous sommes en mesure de suivre la surface oculaire en 3D, de compenser le mouvement de l'œil et de visualiser l'œil entier. La source lumineuse fixée à la caméra constitue un défi et une perturbation lors de l'apprentissage du mouvement de l'observateur. Ce problème a été résolu par la segmentation sémantique et l'ajout d'un nouveau signal de supervision : la fonction de coût de reconstruction sémantique. Nous avons également utilisé la forme de l'œil comme une connaissance *a priori* que nous pouvons inclure comme une contrainte. Ceci a été mis en œuvre par une fonction de coût d'ajustement de forme ; les formes étant deux sphères se croisant l'une l'autre. Notre recalage a montré une amélioration quantitative et qualitative suite à chaque contribution. Nous avons également calculé la concordance inter-observateur des annotations de la kératite ponctuée (zones endommagées). Notre méthode est celle qui se rapproche le plus du niveau d'erreur humaine. La méthode de recalage proposée a également été utilisée

pour une tâche de prétraitement : la sélection des images à analyser. Une fois appliquée à la classification automatique du score d'Oxford, notre méthode a également permis une amélioration des résultats. Cette amélioration valide le fait que les fortes variations de couleur et d'illumination présentes dans les examens constituent une perturbation pour toute tâche d'apprentissage profond. Nous avons surmonté ce problème dans les deux tâches grâce à nos contributions et à la méthode proposée.

TABLE OF CONTENTS

Introduction	13
Context	13
Motivation	14
Thesis Outline	14
1 Clinical Background	17
1.1 Dry Eye Disease	18
1.2 Primary Sjögren’s Syndrome	19
1.3 Clinical diagnostic methods	20
1.3.1 Tear Secretion & Volume	20
1.3.2 Ocular Surface Damage	21
1.3.3 Tear film stability	23
1.3.4 Meibomian Gland Dysfunction	24
1.3.5 Other diagnostic methods	24
1.4 Diagnosis problematics	25
1.5 Conclusion	25
2 Methodological Background	27
2.1 Introduction	28
2.1.1 Deep Learning	28
2.1.2 Projective Geometry	32
2.1.3 Classification: Oxford grade	33
2.2 Deep Learning-based Dry eye quantification	36
2.3 Odometry	38
2.3.1 Visual Odometry	39
2.3.2 Visual-SLAM	43
2.3.3 Structure from Motion	47
2.3.4 Conclusion	49
2.4 Visual Odometry with Deep Learning	49

TABLE OF CONTENTS

2.4.1	Unsupervised deep learning-based methods	52
2.5	Conclusion	53
3	Materials	55
3.1	Introduction	56
3.2	Acquisition Method	56
3.2.1	Original Database 'O'	56
3.2.2	PEPSS Database 'P'	57
3.3	Camera Calibration	59
3.4	Dataset	60
3.5	Ground truth annotations : PEPSS Database 'P'	61
3.6	Conclusion	64
4	Baseline Methods	65
4.1	Introduction	66
4.1.1	Depth prediction for ego-motion	67
4.1.2	Joint depth & egomotion prediction	67
4.2	Conclusion	72
5	Proposed Method	73
5.1	Introduction	74
5.2	Semantic Segmentation	74
5.3	Baseline assessment	78
5.3.1	Inverse Warping	80
5.3.2	Losses	81
5.3.3	Experiments & Results	84
5.4	Semantic reconstruction loss	87
5.4.1	Training	89
5.4.2	Results	90
5.5	Shape Fitting	94
5.5.1	Sphere fitting loss	95
5.5.2	Training	98
5.5.3	Results	98
5.6	Conclusion	100

6 Oxford Grade classification using SiGMoid	101
6.1 Introduction	102
6.1.1 Pre-processing	103
6.2 Evaluation metrics	104
6.3 Multi-class Classification	104
6.3.1 Classification Methods	104
6.3.2 Experiments & results	105
6.4 Binary classification	106
6.4.1 Experiments & Results	107
6.5 Conclusion	113
Conclusion & Perspectives	114
Publications	119
Appendix	120

LIST OF FIGURES

1.1	Annual Cost (2003/2004) of Ophthalmologists in France managing 1,000 Dry Eye Patients	19
1.2	Necessity Working Packages (WPs) [19]	20
1.3	The Oxford grading scale [5].	23
2.1	Artificial Intelligence subsets [81].	28
2.2	CNN training visual example [85].	29
2.3	General CNN architecture pipeline [86]	30
2.4	Learning methods.	31
2.5	Supervised versus unsupervised learning	31
2.6	Camera coordinates	33
	(a) Pinhole camera [94]	33
	(b) World coordinates to camera coordinates [94]	33
2.7	The Oxford grading scale [5].	34
2.8	Classification pipeline [95].	34
2.9	CNN-SPK measurement framework [101].	37
2.10	CNN Tear film break-up time (CNN-BUT) measurement [105].	37
2.11	Deep learning-based Meibomian gland dysfunction diagnosis methods.	38
2.12	SfM vs. V-SLAM vs. VO [109].	39
2.13	ORB vs SIFT vs SURF feature extraction methods [119].	41
2.14	Feature based vs Direct method pipelines [125].	42
2.15	Visual SLAM general pipeline [133].	43
2.16	Visual SLAM algorithms timeline [133].	43
2.17	MonoSLAM [133].	44
2.18	PTAM SLAM [136].	44
2.19	DTAM [121].	45
2.20	SVO [130].	45
2.21	LSD-SLAM [125].	45
2.22	ORB-SLAM 2 [116].	46

2.23	CNN-SLAM [137].	46
2.24	DSO-SLAM [122].	47
2.25	SfM [138].	48
2.26	Geometry-based VO & Deep learning-based VO [160].	50
2.27	Kitti dataset samples [162].	51
2.28	Cityscape dataset samples [177].	51
2.29	Warping process for view reconstruction in unsupervised methods [174–176].	52
2.30	Unsupervised learning of depth and egomotion pipeline [176].	53
3.1	Oxford scale [5].	58
3.2	Examples of DED grading.	58
3.3	Acquisition method.	58
3.4	Haag-Streit image exposure guide.	59
3.5	9 x 10 checkerboard photos through Haag Streit BQ 900 slit lamp.	60
3.6	8 x 7 checkerboard photos through Haag Streit BQ 900 slit lamp.	60
3.7	Reprojection Error	61
3.8	Databases box-plots	62
	(a) Databases vs. Video Length(sec)	62
	(b) Databases vs. No. of frames	62
3.9	Diagnostic test results description.	62
3.10	Labellng tool [179].	63
4.1	Depth estimations from Alexandre Guerre’s thesis [180]	66
4.2	Depth estimation methods	68
	(c) Dense Depth [184]	68
	(d) T^2Net [185]	68
	(e) Iris Depth [186]	68
4.3	Deep learning SLAM methods.	69
5.1	PixelAnnotationTool [190].	75
5.2	Frame augmentation examples.	78
5.3	Differentiable depth image-based warping [176]	81
5.4	Processed frames with binary mask ROI.	88
5.5	Registration evaluation.	90
	(d) Cornea example	90

LIST OF FIGURES

(e)	Sclera example	90
5.6	Visualisation of the \mathcal{L}_{SRL}	93
(a)	Difference between source & target	93
(b)	Difference between registered mask ROI & target	93
5.7	Iterative Closest Point (ICP) algorithm [200].	94
5.8	Point cloud plot.	95
5.9	Human eye anatomy [202].	95
5.10	Modelling of the eye as two intersecting spheres.	96
5.11	Point cloud plots.	99
(a)	Point without \mathcal{L}_{SFL} , Exp. No. C2.	99
(b)	Point cloud using \mathcal{L}_{SFL} , Exp. No. D4.	99
5.12	SiGMoid: Semantic & geometric monocular visual odometry framework. . .	100
6.1	Oxford score [54].	102
6.2	Kalman filtering with left eye video [205].	103
(a)	$m_{cornea} \times frame$	103
(b)	m_{cornea} with center prediction.	103
(c)	Temporal $m_{conjunctiva}$	103
(d)	Nasal $m_{conjunctiva}$	103
6.3	Classical multi-class classification voting method evaluation per patient. . .	105
6.4	Hyperparameters percentage importance for the classification results. . . .	110
6.5	Scatter plot with both modes having select best = False.	111
6.6	Boxplot with both modes having select best = False.	111
6.7	Scatter plot with SiGMoid select best = True.	112
6.8	Boxplot with only SiGMoid select best = True.	112

LIST OF TABLES

1.1	Cost of managing cohort of 1,000 patients	18
1.2	Schirmer’s test and its variations.	22
1.3	Ocular surface staining grading scales	23
2.1	Visual-SLAM methods.	47
3.1	Camera calibration results	61
3.2	Details of Databases	63
3.3	Grader error results.	64
4.1	Depth map predictions part I.	70
4.2	Depth map predictions part II.	71
5.1	Semantic segmentation training examples.	76
5.2	Dice score of all experiments.	77
5.3	Specification of DispNet architecture [164].	79
5.4	Specification of FlowNet architecture [195].	80
5.5	Training input examples.	86
5.6	Experiment loss & training details.	87
5.7	Experiment results.	87
5.8	Experiment details and results A.	91
5.9	Experiment details and results B.	92
5.10	Experiment details and results C.	92
5.11	Depth predictions for experiments.	93
5.12	Experiment details and results D.	99
6.1	Database Oxford score grading.	102
6.2	Validation set classical multi-class classification results.	106
6.3	Test set classical multi-class classification results.	106
6.4	Data split description.	107
6.5	Classification input examples.	108

LIST OF TABLES

6.6	Classification input examples II	109
6.7	Best configuration test set results.	110
6.8	Detailed classification Test set experiment results using S1 data split - Part I.	121
6.9	Detailed classification Test set experiment results using S1 data split - Part II.	122

INTRODUCTION

“Our intelligence is what makes us human, and AI is an extension of that quality.”

— *Yann LeCun*

Context

Due to the evolution of our environment and our lifestyle, more and more people are suffering from dry eyes. According to a recent study, this syndrome affects nearly 7% of the American adult population. Among those with dry eye, some patients have an autoimmune disease called Sjögren’s syndrome, which is characterized by inflammation of the salivary and lacrimal glands. In these patients, dry eye can be severe and requires symptomatic or specific treatments. However, no disease-modifying therapy has been proven to be effective in modifying the course of the disease and preventing ocular damage.

The measurement of ocular dryness is mostly based on an examination of the surface of the eye (cornea and sclera) using a slit lamp (or biomicroscope). The ophthalmologist first applies a contrast medium to the cornea and then observes several signs of dryness. First, he measures the time it takes for the tear film to tear following a blink of the eye, a short time indicating a thin film. Second, it detects the areas of the ocular surface damaged by dryness. These areas appear as dots, specks or filaments and depending on their number and location, a degree of ocular dryness can be determined. The problem with these measurements is that they are not very accurate or reproducible. These limitations prevent a reliable quantification of the evolution of dry eye in a patient. In particular, they do not allow the effect of a treatment to be measured satisfactorily. Our objective is therefore to set up an artificial intelligence designed to perform these measurements in an accurate and reproducible manner.

The objectives of this thesis are multiple and concern the automated processing of videos of the anterior segment of the eye using convolutional neural networks. A database of videos was acquired specifically for this work, during clinical studies including patients followed for a primary Sjögren’s syndrome within the framework of the European project

IMI2 NECESSITY. On these videos, the study of the quality of the blink and the determination of the tear time of the tear film will be the first objective. A classification taking into account the temporal context could be considered (RNN). It will then be necessary to propose a solution to automatically calculate the "Oxford score" corresponding to the location and density of ocular dryness injuries. This second part will require a recalibration between the different images of the video stream, an automatic determination of the appropriate moment for the calculation of the score and a segmentation of the visible lesions.

Motivation

Dry eye disease (DED) is a condition that affects the ocular surface and tear film, resulting in damage. The reason of the visual disturbance can be traced back to a range of medical disorders. The International Dry Eye WorkShop (DEWS) summarizes all of the findings with the goal of bringing existing dry eye disease knowledge up to date [2]. The most recent report creates a DED classification system. The aqueous-deficient, evaporative, or a combination of both causes the tear film to lose homeostasis, or equilibrium, as seen in this disease. DED is a global eye disorder, yet diagnostic approaches are still intrusive, and some grading is non-reproducible.

The main goal of the thesis is to automate a quantification method in order to obtain a more accurate form of DED grading. We focused on providing a complete visual of the eye to help render the process more reproducible. Using artificial intelligence (AI) to its full potential given the scope of the positive results in other fields [6–8]. The following thesis is a *cotutelle* between LaTIM (Laboratoire de traitement de l'information médicale) and LBAI (Lymphocytes B et Autoimmunité). Also involved in a large European project, NECESSITY <https://www.necessity-h2020.eu>, where one of the goals is to automate the quantification of eye dryness.

Thesis Outline

The organization of this manuscript is as follows:

We first take a look at the clinical background in Chapter 1. We detail how DED stems from SS and the current diagnostic methods utilised clinically. Our clinical background research includes four main methods of diagnosis that we prioritised and includes a section that summarises other diagnostic methods. This outline is similar to our review article that not only looks at

existing but also at semi-automated, and fully automated methods. With this article we were able to validate that more methods have emerged in the last couple of years that incorporate automation and deep learning to evaluate and quantify DED.

In the methodological background, Chapter 2, we introduce the framework we wanted to follow to help improve DED quantification. We present the methods we want to utilise, deep learning, projective geometry and lastly classification to predict the DED grade. Following this we present concepts that focus on Odometry, Visual odometry methods with deep learning and more importantly unsupervised methods. We follow this with a description of the materials in Chapter 3. Presenting data we utilised as well as the needed camera calibration and annotations. We extend some of the methods we wanted to utilise as baselines in our work, in Chapter 4. We investigate a few approaches and their essential components in line with the estimation of depth and camera motion prediction. Additionally, we obtained a few poor qualitative results that we wished to further explain.

Chapter 5 first focuses on the development of a baseline method with the use of our evaluation metric. We make an effort to completely comprehend both the reasons why the baselines failed and how they correlate to the challenges in our problematic. To test our theories, we begin by enhancing the primary supervision signals by first adding a more significant loss than the photometric loss. We continue with our more unique loss that emphasizes the use of prior information and shape fitting and achieve better results. We move on to the prediction of the DED grade through classification in Chapter 6. Alongside work completed in an internship, we are able to investigate classical classification. We then incorporate our proposed method to improve the DED prediction.

We finalize by reviewing the overall scope of the work proposed, carried out, and its obstacles. Additionally we consider the several directions and perspectives that could be taken.

CLINICAL BACKGROUND

“Wherever the art of Medicine is loved, there is also a love of Humanity.”

— *Hippocrates*

1.1	Dry Eye Disease	18
1.2	Primary Sjögren’s Syndrome	19
1.3	Clinical diagnostic methods	20
1.3.1	Tear Secretion & Volume	20
1.3.2	Ocular Surface Damage	21
1.3.3	Tear film stability	23
1.3.4	Meibomian Gland Dysfunction	24
1.3.5	Other diagnostic methods	24
1.4	Diagnosis problematics	25
1.5	Conclusion	25

This chapter focuses on the definition, impact and the current state of the art of dry eye disease diagnosis. We look at current clinical diagnostic techniques as well as the rise of automation in the field. Finally, we point out the main issues in the current approaches that we will address in this work.

1.1 Dry Eye Disease

The international dry eye workshop updated definition of dry eye disease to the following [9, 10]:

Dry eye is a multifactorial disease of the tears and ocular surface that results in symptoms of discomfort, visual disturbance, and tear film instability with potential damage to the ocular surface. It is accompanied by increased osmolarity of the tear film and inflammation of the ocular surface.

With a prevalence of 5-50%, the disease impacts include discomfort, visual function and general quality of life . Representing almost 25% of the reasons for consultations in ophthalmology [11]. A study estimated the annual cost associated with the management of DED in six countries summarised in the table below [12]. The study notes that DED prevalence is difficult to measure because it is multifactorial with different definitions and sparse research. The lack of standardization in diagnostic tests also adds to the difficulty to conduct such analyses. It is estimated that the prevalence is expected to increase within the next 40 years [13]. A common cause stems from the fact that we rely daily on computer screens and smartphones, which are reported to cause 30-50% reduction in blinking and therefore increasing the risk of DED.

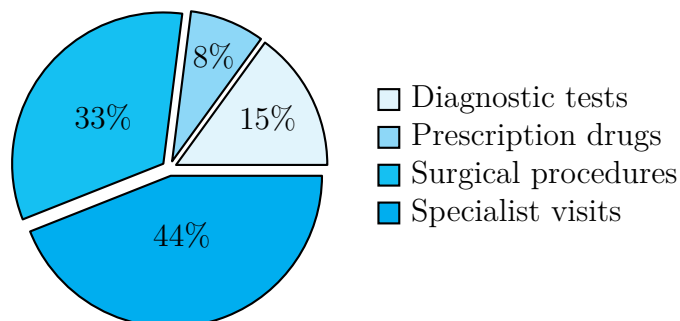
Table 1.1 – Cost of managing cohort of 1,000 patients

Country	US\$ million
France	0.20 - 0.38
Germany	0.41 - 0.66
Italy	0.47 - 0.88
Spain	0.60 - 1.01
Sweden	0.28 - 0.58
UK	0.70 - 1.50

A recent cross-sectional study found that certain etiological subtypes of DED can be found certain demographic and lifestyle factors [14].

DED can be classified into aqueous-deficient and evaporative. Evaporative DED is due to a high evaporative rate of the tear film, caused by Meibomian gland dysfunction (MGD) and lipid insufficiency. The changes in the components of the tear cause instability and goblet cell loss that are responsible for mucins production [15]. Reduced tear production and lacrimal gland dysfunction, characterize aqueous deficient DED. Advancing age, stress and poorer health status were found to be associated to both subtypes. The association of gender was studied and showed that females were more at risk of aqueous deficient DED. Risk factors for evaporative dry eye were found to be contact lens wear, increased screen exposure, stress, age and east and south

Figure 1.1 – Annual Cost (2003/2004) of Ophthalmologists in France managing 1,000 Dry Eye Patients



Total : US \$ 273,000

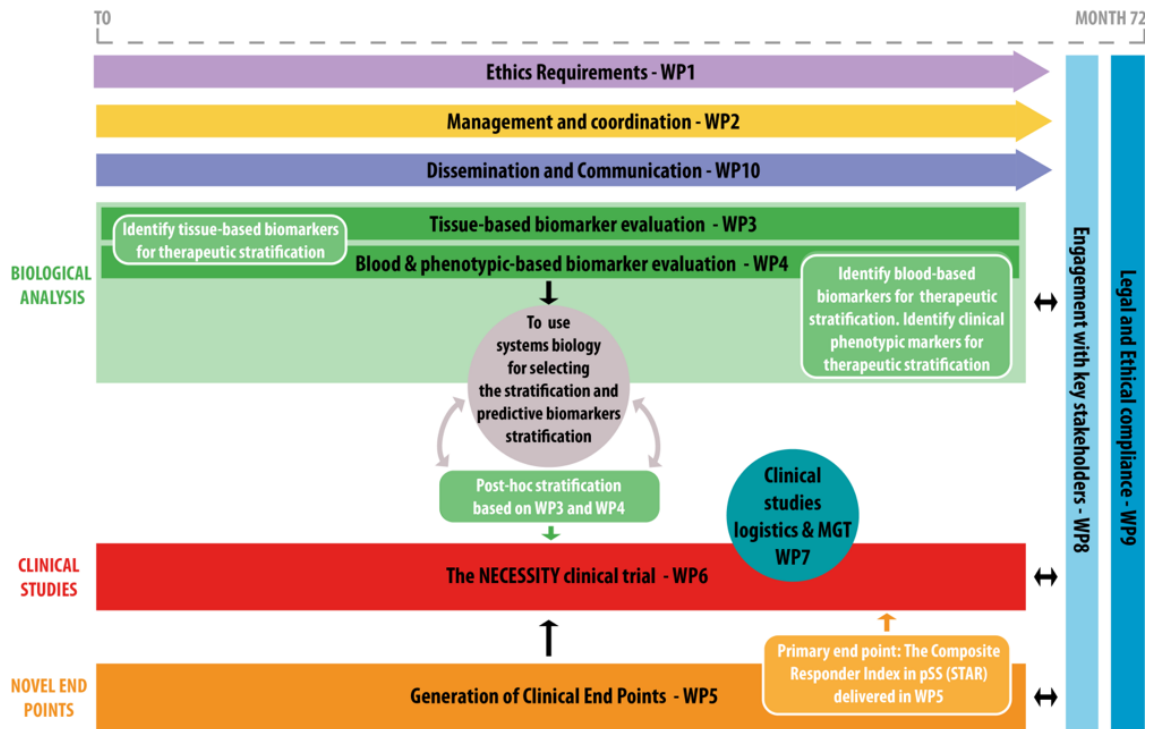
Asian ethnicity. There is a higher population prevalence for evaporative disease caused by MGD or contact lens wear [2, 16, 17]. Symptoms including persistent unpleasant gravel sensation in the eye and the use of tear substitutes.

1.2 Primary Sjögren's Syndrome

Primary Sjögren Syndrome (pSS) is an autoimmune illness that causes dry eyes by targeting the lacrimal glands, mucous membranes, and moisture secreting glands. Various ocular surface diseases can co-exist with dry eye as well, but Dry eye disease (DED) can also be present as a separate condition as-well. One of the main causes of DED is pSS, which is an autoimmune disease that affects the lacrimal glands and results in DED, though not present in all pSS patients [18]. Ten percent of patients diagnosed with DED also have pSS, this further complicates their detection as both have proven to be difficult to diagnose.

NECESSITY Project The aim of the NECESSITY European project is to help identify measures that can be incorporated in clinical trials to test new pSS medicine. The goal is to help solidify clinical trials and quantify their efficacy to bring forward better developed treatments. The primary objective is aimed at evaluating drug treatments for high burden pSS patients. The second objective is to evaluate biomarkers for pSS stratification for organ involvement and disease progression, and the third and last objective is to set-up and execute clinical trials to help validate biomarkers, or newly identified clinical endpoints. With several work-packages as shown in the Figure 1.2 below, the following work takes part in the package title "Novel endpoints: Generation of Clinical End Points - WP5". WP5 also has multiple objectives and ours help develop secondary endpoints that will be innovative tools to better capture the disease progression.

Figure 1.2 – Necessity Working Packages (WPs) [19]



1.3 Clinical diagnostic methods

Many clinical diagnostic methods have been developed over the past decades. These methods have been reviewed and categorized in recent surveys [20–27]. However, we noticed a lack of surveys that take into account the path of automation which was the main outline of our review article [4]. There are clinically used methods as well as methods that have been enhanced by using new equipment or automating the grading alongside new equipment. Given that lack of surveys we decided to review classical, semi-automated and automated methods that aim to diagnose and quantify eye dryness.

1.3.1 Tear Secretion & Volume

One way to quantify DED is to evaluate the decrease in tear secretion and volume. Dryness of the ocular surface appears through different signs and clinical tests that measure the tear meniscus shape and regularity. Reproducible tests that measure secretion and residual volume of tears are key indicators for dry eye. This is because 90% of the tear volume is found in either the superior or inferior tear menisci.

A primary indicator of dry eye syndrome, is the reduction of tear secretion. A deficiency

in any of the layers, ultimately causes discomfort and disrupts the tear film. Estimating tear production precisely allows clinicians to follow through with a suitable course of treatment. Assessing tear secretion dates back to 1903 when Schirmer [28] first presented his test. The test determines whether or not enough tears are being produced to maintain moisture in the eye. It measures both basal and reflex tears [21]. The test uses a small piece of filter paper measuring 35x5 mm that is placed over the lower eyelid. Timing five minutes and measuring the length of the wetted filter paper gives the tear secretion grade. Schirmer describes variations of the test including using a topical anaesthetic, and nasal stimulation to measure reflex tears. Despite the controversy and lack of reproducibility, sensitivity and specificity [22, 29, 30], the test is still used frequently. The use of anaesthetic was investigated, as tear secretion was thought to decrease following its use, therefore, causing misclassification of the damaged ocular surface in staining tests if they are performed afterwards [21]. Limitations of the test include the testing time being too long, the discomfort it causes, and the lack of strict procedure regarding the placement of the paper strip. Modifications have been made to reduce the limitations without major improvements [29, 31–40]. Variants of Schirmer’s test are listed in Table 1.2. Given the limitations and lack of repeatability [41], Nelson *et al.* [42] and Bawazeer *et al.* [43] both describe methods that shorten the procedure to one-minute tests in order to alleviate estimation, without and with topical anaesthetic respectively. Regarding the discomfort aspect, the filter paper was replaced by Kurihasni [44] with a fine thread that was stained with fluorescein at one end. This produced the ‘Phenol red thread test’ [45] where the wetted portion of the thread turns yellow due to the pH of the tears and the length is then measured. Further modifications to the Schirmer’s test include the Fluorescein Clearance Test (FCT) that assesses the tear clearance or turnover rate. These values indicate both the tear secretion and drainage. Clearance is described as normal if the fluorescein dye is no longer detected after 20-minutes. Fluorescein clearance test consists of 1-minute Schirmer’s tests performed consecutively for 30-minutes after the application of the dye [46, 47]. Tear function index value was proposed by Xu *et al.* [48] and consisted of the Schirmer’s test as well as measuring the Tear Clearance Rate (TCR). It is the rate at which the dye fades five minutes after instillation and is graded 1, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64, 1/128 and 1/256. The tear interference device allows to non-invasively visualise the tear meniscus, first reported by Guillon *et al.* [49]. Strip meniscometry was investigated by Dogru *et al.* [50] that eliminates the use of fluorescein dye, or any touching of the eyelid and is performed in 5s.

1.3.2 Ocular Surface Damage

The ocular surface provides both anatomic and immunologic protection as well as functions to maintain clarity for the cornea [51]. Given the sensitivity of the structures that it helps and protects, any damage to the ocular surface can produce severe consequences. Instillation of a

TABLE 1.2 Schirmer’s test and its variations.

1903	Schirmer Test (Schirmer 1903)
1982	Short Schirmer I (Nelson 1982)
1983	Phenol red thread (Hamano et al. 1983)
1995	Tear Function Index (Xu et al. 1995)
1998	Fluorescein Clearance test (Prabhasawat & Tseng 1998)
1999	Fluorescein Clearance test (Fluorophotometry) (Afonso et al. 1999)
2003	Short Schirmer I (with anesthesia) (Bawazeer & Hodge 2003)

dye causes the penetration of the lipid layer of the epithelium, and staining areas are where shed cells are highlighted. These areas show where the epithelium has been damaged. Although not a specific sign of dry eye, staining can quantify the damage done and its severity. The limitations of visual scoring and grading of ocular surface damage motivate the need for improvement.

Damage to the corneal epithelial is stained and made more visible using fluorescein sodium. Instilled using paper strip or preserved doses, the staining is more visible if a yellow (blue-free) filter is placed in the slit-lamp [22]. Damage to the conjunctival epithelium is more difficult to detect with fluorescein staining due to the poor scleral contrast. One examination is also not sufficient to evaluate the damaged ocular surface when using ocular staining. Another dye which is a derivative of fluorescein is Rose Bengal (RB), which is mainly used on the conjunctiva to detect damage to the epithelium [22]. RB staining was shown to stain areas that lack membrane-associated mucins [52]. Lastly, Lissamine green is a synthetic organic acid dye that is interchangeable with RB [25], though used more often since it has been proven to be less toxic and more easily tolerated.

Dating back to 1882 when fluorescein was used to stain corneal abrasions [53]. Bron *et al.* present a more exhaustive coverage of clinical ocular surface staining [54]. Numerous grading scales for ocular staining have been developed as detailed by Begley *et al.* [27]. Grading scales mentioned in Table 3.3 were studied on dry eye subjects. Another more detailed method of grading recently developed by Woods *et al.* [55] includes a grading scale of 0-100 for staining and extent (area) and 0-4 for depth. Named the CORE (Centre for Ocular Research & Education) staining scale, it can also be reported for each of the five zones (central (C), superior (S), nasal (N), inferior (I) and temporal (T)) which could aid in tracking the evolution of the damage.

The Oxford scale [5] includes six grades (0-5) with the dots ordered on a log scale. Figure 2.7 is the Oxford scale that is referred to visually after the examination to determine the patient’s grade. First a dye is instilled that stains the dryness on the ocular surface. The damage is then made more visible and in order to quantify it a grade can be obtained by referring to a scale during the examination. By examining the eye through a slit-lamp using standard settings the

examiner starts to grade the eye, which for the Oxford score is decomposed into three panels. The three panels being: nasal conjunctiva, the side adjacent to the nose, the temporal that is adjacent to the temple for each of the eyes and lastly the cornea in the middle. For the examination the upper eyelid is raised to have a better complete visual of the cornea. Each of the panels, as shown in the figure are graded from 0-5 and the final Oxford score is the sum and ranges from 0-15. The grading scale is an estimated number of dots, that are of course impossible to count during the examination, and the log of the dot numbers is the final grade.

Figure 1.3 – The Oxford grading scale [5].





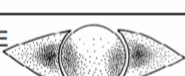
PANEL	GRADE	CRITERIA
	0	Equal to or less than panel A
	I	Equal to or less than panel B, greater than A
	II	Equal to or less than panel C, greater than B
	III	Equal to or less than panel D, greater than C
	IV	Equal to or less than panel E, greater than D
>E	V	Greater than panel E

Table 1.3 – Ocular surface staining grading scales

Scale	Year	Zone		Grading description
		Cornea	Conjunctiva	
Van Bijsterveld and Utrecht (Bijsterveld 1969)	1969	Whole	2 zones	0-3 each zone based on staining intensity
Lemp (Lemp & Michael 1995) (NEI/Industry Workshopscale)	1995	5 zones	6 zones	0-3 each zone based on staining intensity
Bron <i>et al.</i> (Bron 1997, Bron, Evans & Smith 2003) (Oxford Score)	1997	Whole	2 zones	0-5 each zone (log-linear increase)
Mitaya <i>et al.</i> (Miyata 2003)	2003	Whole	N/A	Combination score
De paiva <i>et al.</i> (De Paiva & Pflugfelder 2004) (Baylor Scale)	2004	5 zones	N/A	0-40 sum of corneal zones
Whitcher <i>et al.</i> (Whitcher et al. 2010) (SICCA OSS group scale)	2010	Whole	2 zones	0-3 each zone based on counting dots
Abelson <i>et al.</i> (Abelson et al. 2016) (ORA Calibra Scale)	2016	3 zones	2 zones	0-4 grade (none to severe staining)
Woods <i>et al.</i> (Woods et al. 2018) (CORE)	2018	5 zones	2 zones	0-100 grade (area)

1.3.3 Tear film stability

The priority in assessing the tear film is to be non-invasive yet able to represent its instability. The quality of each layer within the tear film is important for its stability, accordingly, measuring

it is essential to characterize DED. The most referenced and useful technique to assess the extent of tear evaporation is the tear break-up time (TBUT). Introduced by Norn *et al.* [56] the test aims to diagnose the tear films instability by instilling sodium fluorescein, and timing between the last blink and appearance of the first break or dry spot: less than 10s is abnormal, 5-10s being marginal and less than 5s suggests dry eye. By observing through a slit-lamp, the first appearance of a dry spot or a tear in the film indicates the TBUT. TBUT is performed in various ways and is being continuously modified. The main difference between the ways the test is performed is the degree of invasiveness. Current methods include instilling sodium fluorescein, and observed using a cobalt blue light with a yellow filter. On the other hand, non-invasive techniques do not involve instilling dye and instead use different diagnostic instruments. The temperature change mapped using an ocular thermogram allowed Morgan *et al.* [57] to determine that the mean ocular surface temperature was greater in dry eye patients. Fujishima *et al.* [58] determined a change in the corneal temperature using an infrared radiation thermometer. Changes in temperature with each blink was observed to be smaller in patients with dry eye. Another instrument is the keratometer which measures the corneal curvature. Observation mostly includes an illuminated grid pattern reflected from the tear surface [24]. A modified method of keratometer, proposed by Hirji *et al.* [59], includes adding a circular grid and the mean of five measurements to obtain the TBUT.

1.3.4 Meibomian Gland Dysfunction

Meibomian gland dysfunction (MGD) occurs when the Meibomian glands become blocked, resulting in a lack of oil production, disrupting homeostasis and causing the tear film to evaporate too rapidly. Tear Film and Ocular Surface Society (TFOS) completed a report on MGD in 2011, a leading cause of DED [60]. MGD treatments are often based on severity which only underlines the need for a precise diagnostic method. Xiao *et al.* [61] recently addressed this through a study where patients were classified into dry eye severity level (1-4) [62]. The study compares morphologic and functional parameters including length, density, thickness and quality from Meibography images. Results found that Meibograde, gland distortion and MG length differentiate between DED and non DED.

1.3.5 Other diagnostic methods

Different approaches that detect symptoms of dry eye can also be used to quantify the severity of the disease. A very simple method that is employed clinically is questionnaires: Dry Eye questionnaire, McMonnies questionnaire, and Ocular Surface Disease Index (OSDI) are assessed by Simpson *et al.* [23]. The study shows that responses from the three questionnaires

met the Rasch analysis criterion of unidimensionality, where the Rasch analysis represents a certain structure for a data that enables a successful measurement. A detailed review of existing questionnaires in Japan providing their contents and characteristics is assessed by Shiraishi *et al.* [63]. Several other reviews assess dry eye questionnaires including [64–68]. Another method is to estimate tear osmolarity and protein concentration in tear film to quantify dry eye. An osmolarity referent of 316 mOsmol/L found to be a good cutoff value to determine tear hyperosmolarity [69]. Automated osmolarity is measured often with a tear osmometer as presented by Suzuki *et al.* [70], where it correlated with Schirmer’s test ($r=-0.52$) and a microfluidic approach by Karns *et al.* [71].

An automated evaluation has been made possible using the TearLab Osmolarity System (TearLab Corp., San Diego, CA) which facilitates tear sample collection and displays the result of the ion concentrations. A more precise method of analysis of the tear film focuses on the primary component of the external layer, the meibum. Various collection and quantification methods are addressed by Pucker *et al.* [72]. The change in protein expression of tear film proteins was studied by Srinivasan *et al.* [73] in DED and non DED patients. Isobaric tags for relative and absolute quantitation (iTRAQ) was used, followed by protein information resource (PIR) to interpret pathways and protein functions. Differences were detected between the two groups that correlated with Schirmer’s and OSDI scores.

1.4 Diagnosis problematics

Although clinical diagnostic tests vary and there are many ways to detect the disease, there is no standard test to this day [74]. Clinical diagnostic tests are limited in terms of quantifying dry eye, which inherently makes dry eye severity assessment also a challenging task. Measuring the extent of dry eye has also proven to be a difficult task. Most classical methods lack efficiency in certain aspects including repeatability, accuracy and reproducibility. The complexity of the disease causes varied signs, symptoms and extreme changes with seasons, time of day and ultimately eye care examinations [75]. There are also asymptomatic patients that are overlooked, or even physical conditions, such as like floppy eyelid syndrome, lid imbrication syndrome, and conjunctivochalasis, that cause misdiagnosis of DED [76].

1.5 Conclusion

With the evident burden dry eye holds on patients, it is becoming more critical to standardize the diagnosis. It is currently critical to look at patient history and utilize at least two clinical examinations to assess patients. There is a clear motivation for the need of an automation

for DED diagnosis to better evaluate current patient treatment. The review we conducted for classical, semi-automated and fully-automated diagnostic methods displayed a clear trend for the incorporation of artificial intelligence [4]. We move on in the next chapter to detail the key elements we will use in this work.

METHODOLOGICAL BACKGROUND

“AI has by now succeeded in doing essentially everything that requires ‘thinking’ but has failed to do most of what people and animals do ‘without thinking’-that, somehow, is much harder.”

— *Donald Knuth*

2.1	Introduction	28
2.1.1	Deep Learning	28
2.1.2	Projective Geometry	32
2.1.3	Classification: Oxford grade	33
2.2	Deep Learning-based Dry eye quantification	36
2.3	Odometry	38
2.3.1	Visual Odometry	39
2.3.2	Visual-SLAM	43
2.3.3	Structure from Motion	47
2.3.4	Conclusion	49
2.4	Visual Odometry with Deep Learning	49
2.4.1	Unsupervised deep learning-based methods	52
2.5	Conclusion	53

This chapter discusses the main objectives of this research in a detailed manner. We also take a look at the key elements used to investigate a new method for DED quantification. The chapter also details the state of the art of the key elements.

2.1 Introduction

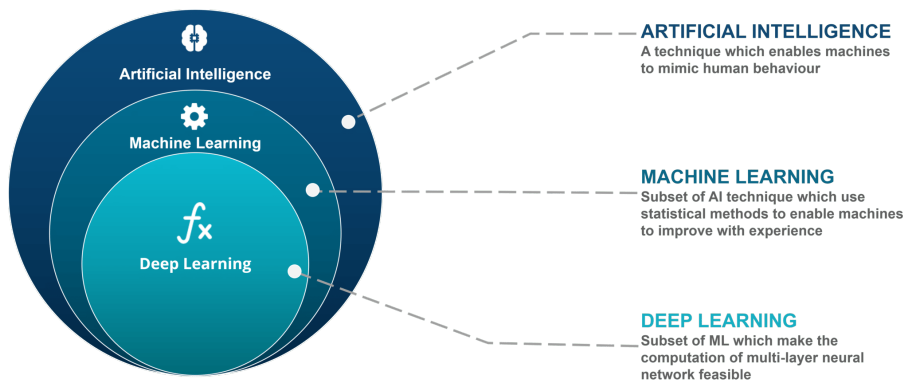
The objective for this research is to utilise the examinations for dry eye diagnosis along with artificial intelligence and obtain a new method to automate the quantification. The dataset is a collection of slit-lamp video examinations of DED along with annotations of the Oxford grade in Figure 2.7. To facilitate the problem we process the frames from these videos to first register them to a common coordinate system. One of the ways we evaluate our proposed solution is by predicting the Oxford grade for each patient. There are three key elements that makeup the following thesis, AI, projective geometry and lastly classification.

2) Projective geometry : image registration } *using* → 1) Artificial Intelligence : Deep learning
 3) Oxford Grading : Classification

2.1.1 Deep Learning

Deep learning (DL), the most popular branch of AI, is already a pillar in automated methods in various fields [6, 7] as well as medical image analysis [8]. DL is a variant of the traditional neural network, however it performs much better than its foundations. Convolutional neural networks (CNNs) are widely used in the field of deep learning. The main advantage of CNNs over classical methods, is that they identify relevant features without any human supervision. They have been applied in a range of different domains, including computer vision, speech processing, facial recognition, etc [77–80].

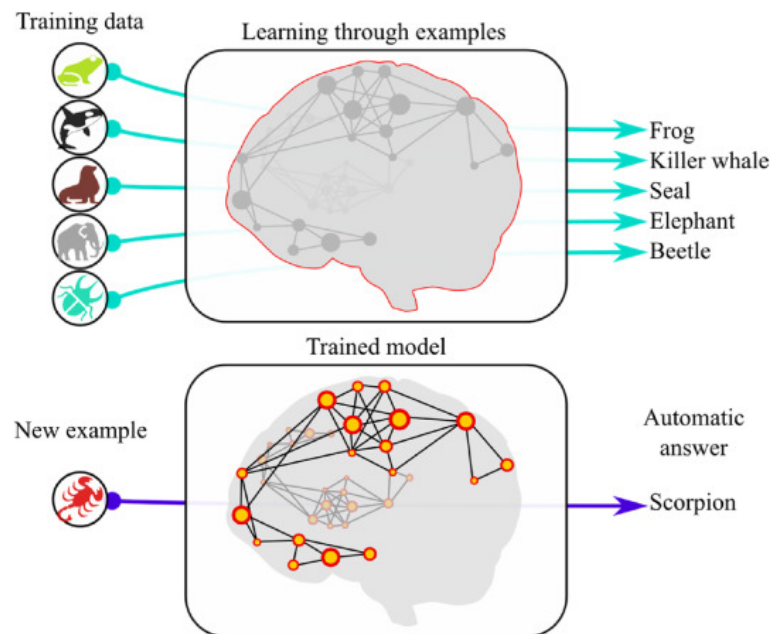
Figure 2.1 – Artificial Intelligence subsets [81].



The global AI in healthcare market is projected to grow almost 46%, reaching USD 95.65 Billion by 2028, up from USD 6.60 Billion in 2021 [82]. Researchers have experimented on CNNs to help in disease detection, recognition and ultimately to optimise interpretation [83]. Some of the applications that have been examined and enhanced in medical imaging using CNNs include:

classification, object detection, image segmentation, image generation and transformation [84]. A simple visualisation of classification training is shown on Fig 2.2. CNNs employ shared weights and local connections, in contrast to multilayer perceptrons (MLP), to fully exploit 2D input data structures like found in images. The training process is sped up, made simpler, and employs a very limited number of parameters in this design.

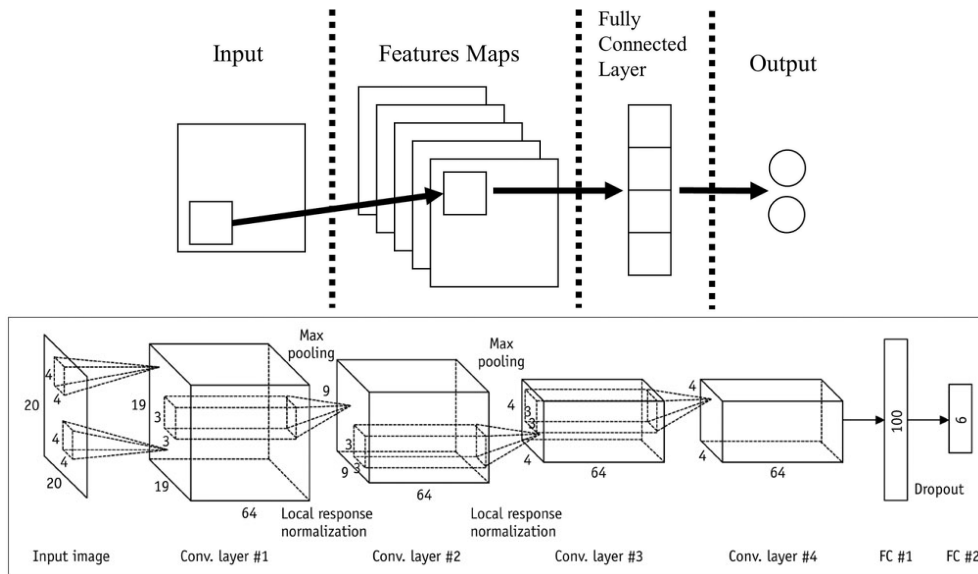
Figure 2.2 – CNN training visual example [85].



Convolutional, pooling, and fully linked layers make up CNN's architecture (Fig.2.3). A convolutional layer's main objective is to find distinct local edges, lines, and other visual components. Convolutions are specialized filter operators whose parameters are learnt. This mathematical process denotes the multiplication of a particular pixel's immediate neighbors by a tiny set of previously learnt parameters known as a kernel. This procedure imitates the extraction of visual characteristics, such as edges and colors, similar to that observed for the visual cortex, by learning relevant kernels. Filter banks can be used to complete this procedure. Each filter can be seen as an item with a square form that moves over the provided input or image [86].

We believe that the application of artificial intelligence (AI) in medical imaging will operate as a collaborative tool for reducing the strain and distraction from numerous mundane and repetitive jobs. This is why our first and key element for this research is to accomplish the objective by employing deep learning techniques.

Figure 2.3 – General CNN architecture pipeline [86]



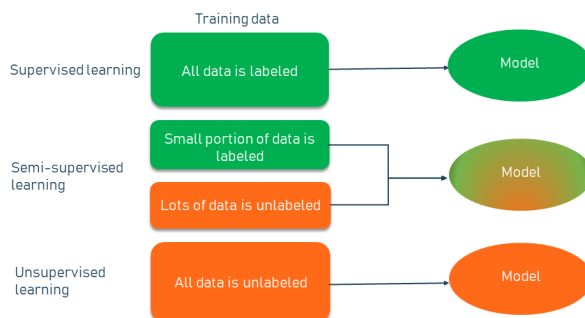
Learning methods

The most common learning method used in AI systems is supervised learning, which involves feeding the learning system "labeled" training data (data samples paired with the relevant class or label), in order to train the model. Finding a relation that converts each input of the training set (the data) into an output is the learning system's goal (the label). In medicine, using various imaging modalities, the output label might be anything from the disease diagnosis to the patient's state (such as the disease stage at a specific follow-up period) to the treatment outcome.

The distinctions between supervised and unsupervised architectures are shown in Figure 2.5. In both approaches, the inputs are fed and then adjust the network weights to reduce errors between the produced output and the predicted result (supervised learning) or between the similarity of the input signals and the output. Unsupervised learning minimizes error by comparing the various inputs, but supervised learning necessitates labels for error optimization. Additional methods can also be utilized, such as semi-supervised learning, which combines supervised and unsupervised learning by labeling only a portion of the training data [87, 88]. Semi-supervised learning is a third method where the input dataset is a mix of labelled and unlabelled inputs. To address their main problems, semi-supervised learning bridges supervised learning and unsupervised learning methods. A self-training method is a simple method of semi supervised learning where initially a model is trained on a small sample of labeled data before expanding it repeatedly to a larger sample of unlabeled data. An improved version is co-training [89], a semi-supervised technique, where two individual models are trained based on two views of the data. This means that the datasets have different sets of features that can stand alone

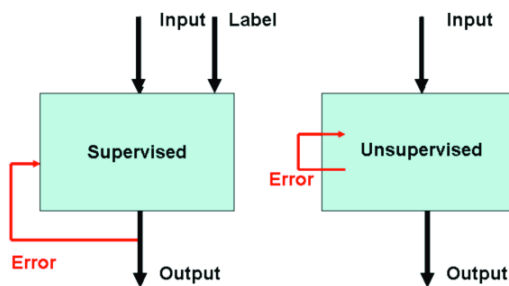
and reliably train a model.

Figure 2.4 – Learning methods.



There are various learning methods, Figure 2.4, and a subcategory of unsupervised learning is self-supervised learning [90]. This refers to using data that doesn't have manually added labels 'pseudo-labels'. Models are trained to learn good representations of objects, this is called a 'pretext task'. A pretext task is one with artificially created labels, or 'pseudo-labels'. Finally a downstream task is a task that evaluates the quality of features learned by self-supervised learning. This is when a model that has been previously trained (self-supervised learning) or simply components are used to perform tasks such as image recognition or segmentation in a supervised learning pipeline.

Figure 2.5 – Supervised versus unsupervised learning .



2.1.2 Projective Geometry

We perceive our three dimensional world in 2-D, and what we perceive is a succession of 2-D projections. In the case of using a camera, as it moves we need to estimate this motion and this is possible through projective geometry. We are also able to model the scene through these 2-D acquisitions. In projective geometry, projective transformation depicts objects from different points of view. We have affine geometry which includes Euclidean geometry. In order to model a camera imaging process, we often refer to projective geometry since it includes a larger class of transformations. Projective transformations also preserve type, incidence and cross ratio but not size nor angles [91]. These are important properties in projective geometry, where incidence is the heterogeneous relation between a point and a line.

In 3D space \mathbb{P}^3 , the homogeneous coordinates of a point is represented by a 4-vector as

$$\mathbf{X} = [x_1, x_2, x_3, x_4]^T$$

which is defined up to a scale since \mathbf{X} and $s\mathbf{X}(s \neq 0)$ represent the same point. A plane in 3D space \mathbb{P}^3 can be formulated as :

$$\Pi\mathbf{X} = \pi_1x_1 + \pi_2x_2 + \pi_3x_3 + \pi_4x_4 = 0 \quad (2.1)$$

Determining three-dimensional information from two-dimensional images is referred to as the reconstruction problem. We want to determine the 3D scene and the camera viewpoints from a series of images [92–94] . The use of one formula allows for a straightforward formulation of this issue. A scene is made up of a collection of 3D points. The projection of the 3D point \mathbf{X} to the image point \mathbf{x} using a camera with a Cartesian center $\bar{\mathbf{Q}}$ and matrix \mathbf{H} is as follows:

$$\mathbf{x} = H(\mathbf{X} - \bar{\mathbf{Q}}) \quad (2.2)$$

A simplification of a real world camera is called a pinhole camera. Where we consider it as a central projection device of the euclidean space \mathcal{E}^3 onto the image plane \mathcal{E}^2 . Mapping the 3D point \mathbf{X} onto a point \mathbf{x} on the image plane is written as :

$$\mathbf{X} = (X, Y, Z, 1)^T \quad \text{onto} \quad \mathbf{x} = (x, y, 1)^T \quad (2.3)$$

$$x = f\frac{X}{Z} \quad \text{and} \quad y = f\frac{Y}{Z} \quad (2.4)$$

The centre of projection is camera centre $\bar{\mathbf{Q}}$, and the focal length f is the distance between the camera centre and the image plane through $\bar{\mathbf{Q}}$. Aspect ratio r and the skew s of a pixel are also parameters of the camera model. The optical axis of the camera is defined as a line that is

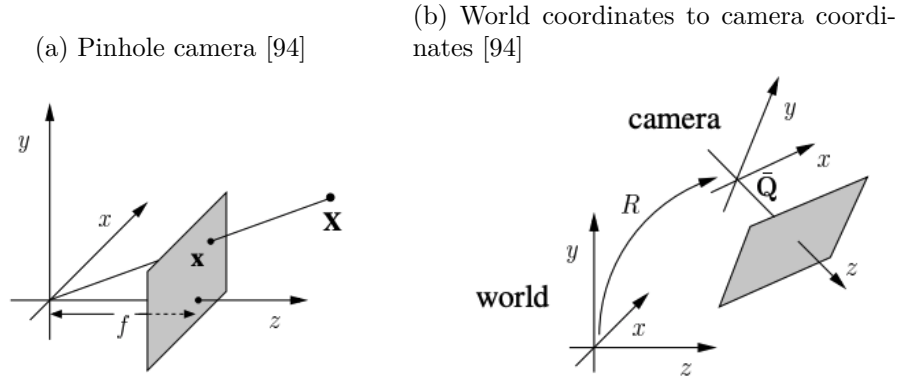
perpendicular to the image plane and passes through its center. The principal point x_0 of the camera is where the optical axis and image plane cross. The pinhole camera is shown in figure 2.6(a) where the principle plane is z . The camera coordinate system can be assumed to align with the world coordinate system, but in general it is rotated and translated with respect to the world coordinate system (Figure 2.6(b)).

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \sim \begin{pmatrix} f & s & x_0 \\ 0 & rf & y_0 \\ 0 & 0 & 1 \end{pmatrix} (I_{3 \times 3} | \mathbf{0}_{3 \times 1}) \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \quad (2.5)$$

$$\mathbf{x} \sim K(I_{3 \times 3} | \mathbf{0}_{3 \times 1}) \mathbf{X} \quad (2.6)$$

$$\mathbf{x} = K[R|t] \mathbf{X}_{world_coord} \sim P \mathbf{X}_{world_coord} \quad (2.7)$$

Figure 2.6 – Camera coordinates



Matrix K is the calibration matrix that has the intrinsic camera parameters, unique to each camera, and constant over time. These parameters are determined by camera calibration, which is later discussed in details in chapter 3. Extrinsic camera parameters are the R and $\bar{\mathbf{Q}}$ that change over time.

2.1.3 Classification: Oxford grade

One of the ways to assess DED, is through the damage caused which is made visible through staining of the ocular surface. Dyes highlight the damaged ocular surface of the conjunctiva and cornea. The damage on the surface, referred to as punctate dots, can be very hard to count, and current grading scales that describe the patients' state are very broad. There were detailed in section 1.3.2. One of the methods we detailed is the Oxford scale [5], which includes six grades

(0-5), where the number of dots are ordered on a log scale. Figure 2.7 shows the Oxford scale that is referred to visually during the examination to determine the patient’s grade.

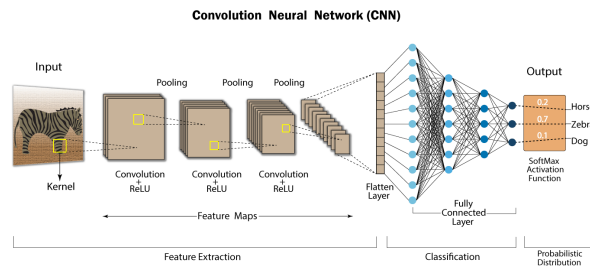
Figure 2.7 – The Oxford grading scale [5].

PANEL	GRADE	CRITERIA
A	0	Equal to or less than panel A
B	I	Equal to or less than panel B, greater than A
C	II	Equal to or less than panel C, greater than B
D	III	Equal to or less than panel D, greater than C
E	IV	Equal to or less than panel E, greater than D
>E	V	Greater than panel E

In order to evaluate our improved method of quantification we implement an Oxford grading classification as an evaluation method. Given the material and annotations we have that are detailed in chapter 3, we are able to evaluate using deep learning methods.

Ultimately to accurately grade the damage to the ocular surface, a full ‘eye map’ is needed. A map where all the punctate dots are visible, to minimize disregarding some or overestimating the damage. Acquiring a full mapping can also help in evaluating a patient’s progress given the medical treatment they are following. The concept of classifying images involves taking an input, such as an image, and producing a class, or a probability that the input belongs to a specific class. As a result, we want to finish by including DED grade classification in our proposed method. Focusing on the Oxford score, we can learn to classify frames from our examinations and predict the severity of the DED. A general classification pipeline is shown below in Figure 2.8.

Figure 2.8 – Classification pipeline [95].



In computer vision, it is one of the most popular application fields. CNNs classify images by

employing a sequence of feature extraction layers followed by a classification layer. The evolution of CNNs has included various upgrades throughout time . Most recently, for COVID-19 screening various classification deep learning models were used [96–100].

In the rest of the chapter we will look at AI solutions that exist for image registration, and classification. First taking a look at the exact problem we want to solve given the assumptions we made:

1. Movement observed is due to camera motion relative to the eye
⇒ multiple view geometry
2. Only data available are the examination videos
⇒ unsupervised learning

Problem :

- Learn to register multiple camera views to one coordinate system in an unsupervised manner.

More specifically, we focus on AI solutions that help learn image registration from multiple camera views in an unsupervised manner. Other AI solutions we look into are image segmentation and classification. Image segmentation is used to identify objects and boundaries (such as lines, curves, etc.) in images. Classification is the process of identifying to which predetermined categories images or pixels pertain to. We begin the methodological background by looking at approaches that aim to quantify DED using AI.

2.2 Deep Learning-based Dry eye quantification

As we mentioned in Chapter 1, our extensive review article includes clinical diagnostic methods, semi-automated and fully-automated methods. For DED diagnosis some of the fully-automated methods were deep learning based. We go into more depth about some of the deep learning approaches we looked at below, since deep learning will be used throughout our work to help improve DED diagnosis.

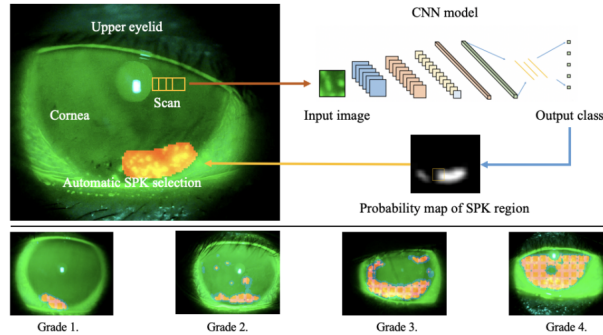
Benefiting from deep learning, Su et al. [101] propose the automatic detection and grading of punctate dots with a Convolutional Neural Network (CNN). Images were manually segmented only to train the model using five pre-defined classes: tear film, eyelash, eyelid, punctate dots and conjunctiva. The model then produces a probability map of punctate dots, used to calculate the CNN-SPK value, a newly defined grading scale, which is:

$$\text{CNN-SPK} = \frac{\text{area of the punctate dot predicted}}{\text{area of the cornea}} \quad (2.8)$$

Obtaining significant correlations with clinical grading scales, this method is close to those presented in [102–104]. Figure 2.9 demonstrates the proposed framework.

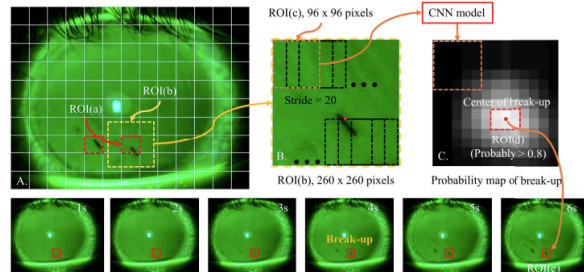
Another advanced study to detect tear break-up time utilizes a deep CNN and was also proposed by Su *et al.*[105]. Digital slit-lamp recordings were used to train the CNN model. The

Figure 2.9 – CNN-SPK measurement framework [101].



method first labels is patches of each frame into : break-up, non-break, eyelash, eyelid, and sclera. The trained model able to identify patches as break region then results in a probability map of break-up, when this exceeds a threshold the tear break-up time is set. This study is the first CNN application to evaluate tear break-up time and it resulted in strong correlations with clinical measurements proving that it could be further improved. The proposed method is also presented in Figure 2.10 below.

Figure 2.10 – CNN Tear film break-up time (CNN-BUT) measurement [105].

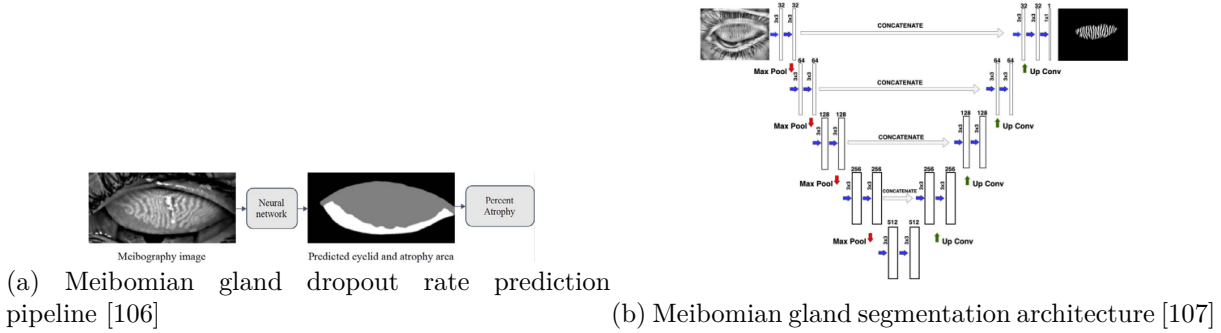


Both methods that are in line with the imaging system we have, cameras attached to slit-lamps, as well as the method of diagnosis we wish to focus on, rely heavily on annotations. Unfortunately, they are both time and data intensive approaches.

Our review article also includes an overview of other diagnosis methods, such as those that quantify Meibomian gland dysfunction. This area also included a few deep learning-based approaches. Fully automated image-based solutions rely on tasks enhanced by deep learning. One of which is automated segmentation, and it has been used to quantify Meibomian gland dropout rate. Wang et al.'s proposed model segments and computes atrophy percentage achieving an accuracy of 97.6% and 95.4%, respectively [106]. Using 706 annotated images the proposed method displayed an accurate evaluation of gland atrophy and ultimately DED diagnosis through Meibomian gland dysfunction. A more recent study by Prabhu *et al.*[107] uses CNNs to segment

Meibomian glands. Comparing the p -value > 0.005 between the ground truth segmentation and various trained models, it showed that the model trained with data augmentation improved accuracy. These automated gland segmentation methods using deep learning pave the way for improvement within the field.

Figure 2.11 – Deep learning-based Meibomian gland dysfunction diagnosis methods.

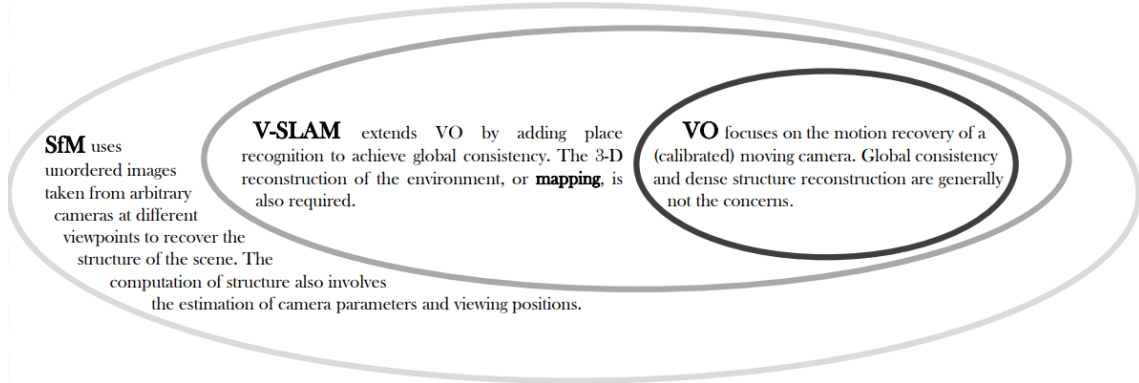


Although we only highlight deep learning-based solutions in this section, our review article [4] of DED diagnosis methods includes other fully automated approaches as well.

2.3 Odometry

Odometry is the process of estimating changes in location over time using data from motion sensors. Some wheeled or four-legged robots in robotics use it to calculate their position in relation to a starting point. We focus on visual odometry where the data inputs are visual elements from associated camera images. Firstly we take a look at the definition and aims of Visual Odometry (VO), Visual Simultaneous Localization and Mapping (VSLAM), and Structure from Motion (SfM), and how they relate to each other. The main general concept is SfM that is an offline method that uses unordered sequences of images and it aims to map the environment. The images are from different perspectives and can even be from different cameras. SLAM’s main goal is a global consistent estimate of a robot’s path. Besides localization, loop detection and loop closure are the main issues in SLAM [108]. Lastly, VO is the key components of VSLAM, and the main process of detecting the orientation and location of the robot. Figure 2.13 shows the relation between the three concepts. We go into more detail on visual odometry first, and the different data inputs and estimation methods.

Figure 2.12 – SfM vs. V-SLAM vs. VO [109].



2.3.1 Visual Odometry

Many applications in computer vision depend on camera ego-motion estimation, also known as Visual Odometry (VO) [110]. Over the years VO has become a viable method for vehicle localisation using a stream of images captured by the camera attached to a vehicle. The applications of VO include not only autonomous driving but also medical robots, augmented and virtual reality. Optical cameras are one of many sensors employed and has been of great interest for visual-based localization [111]. It is also an inexpensive alternative, that is comparatively more accurate [112]. Given the nature of our problem we focused on optical camera vision based methods. Types of camera/data sensor used for VO estimations are:

1. Stereo-camera
 - Also referred to as a binocular camera, it has two lenses which separate image sensors for each lens.
2. Monocular cameras
 - Only a single camera sensor or lens is present.
3. Stereo or monocular omni-directional cameras
 - omni-directional cameras have a very wide field of vision, and therefore more information.
4. RGB-D cameras
 - Both color and dense depth images can be obtained from RGB-D (color-depth) cameras in real time. Depth information is obtained through the 3D depth sensor.

Each of these camera sensors have advantages and disadvantages. There are various VO estimation methods that can be implemented with any of the camera sensors. These are considered as a module for Visual-SLAM, which we will discuss in detail in the coming section.

Feature-based methods

These type of methods rely on extracted image features (corners, lines, curves) and either matching or tracking distinctive features throughout a set of frames. The images are matches by comparing the features and measuring the Euclidean distance of the vectors to match the features. The displacement is then measured by calculating the velocity vector between the matches feature points. Camera motion is also estimated by measuring the relative pose of the camera through geometry transformations between two images. A common computational method to match feature points is by determining nearest neighbour pairs from feature descriptors [110]. The extracted feature points are used to project two-dimensional points. Most VO implementations assume a calibrated camera, or require this to be performed beforehand. Algorithms that detect, describe, and match local features often have a high computational cost. Three main feature point extraction methods are detailed below :

1. Scale-invariant feature transform (SIFT) [113]
2. Speeded-up robust features (SURF) [114]
3. Oriented FAST and rotated BRIEF (ORB) [115, 116] : combines the advantages of features from accelerated segment test (FAST) [117] and binary robust independent elementary features (BRIEF) [118].

Lastly, after detecting and matching features the final step is to calculate relative motion between frame. Depending on the available data one of the following approaches can be employed :

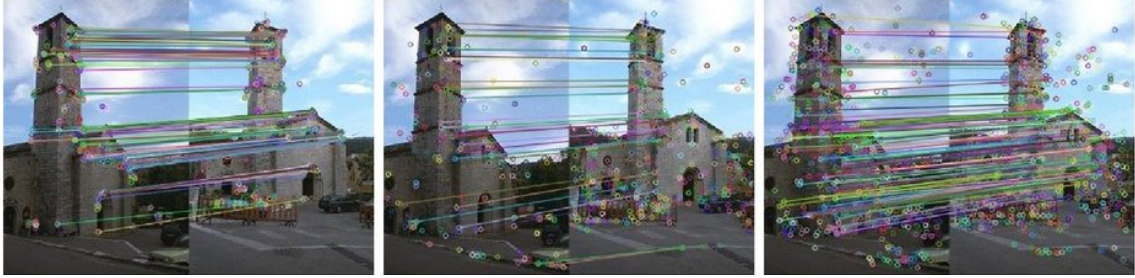
- Perspective-three-point (P3P)
- Iterative closest point (ICP)
- Epipolar geometry

More precision can be achieved through iterative optimization such as bundle adjustment. Bundle adjustment consists of reducing the reprojection error between observed image locations and predicted ones. Through this nonlinear least-squares method we achieve to minimize the error. Outliers are also addressed through the iterative process of Random sample consensus (RANSAC).

Direct tracking based methods

Direct based methods, developed from optical flow, estimates camera motion and pixel spatial location mainly by minimizing the photometric error (reprojection error). This method does not rely on any feature extraction or feature description. Direct methods directly calculate the structure and motion based on the image's intensity data. The size and direction of the gradient are employed for optimization. In terms of robustness with insufficient textures or unfocused cameras, direct techniques perform better than feature-based methods [120]. Direct approaches work directly on the image's intensity values, which speeds up feature detection [121].

Figure 2.13 – ORB vs SIFT vs SURF feature extraction methods [119].



The three main categories are :

- Sparse direct method (DSO): use actual sensor values-light as measurements and optimizes photometric error without incorporating geometric prior [122].
- Dense direct method: use a photometric error as well as a geometric prior to estimate dense or semi-dense geometry. Examples include DTAM[121, 123], REMODE [124].
- Semi-dense direct method: use the connectedness of the used image region to formulate a geometry prior. A method proposed by Engel *et al.*[125] uses direct image alignment coupled with filtering-based estimation of semi-dense depth maps.

Direct tracking based VO approaches were rarely based on the tracking and mapping framework; instead, the majority of it focused on the primary areas of artificial selection [126–128]. Recently, it seems that direct approaches could use the geometry and grayscale information directly from the image pixels to generate the error function through the graph optimization to minimize the cost function, thereby achieving the ideal camera pose. Large-scale map problems with pose graph are addressed using these techniques [122, 129].

Hybrid of feature- and appearance-based methods

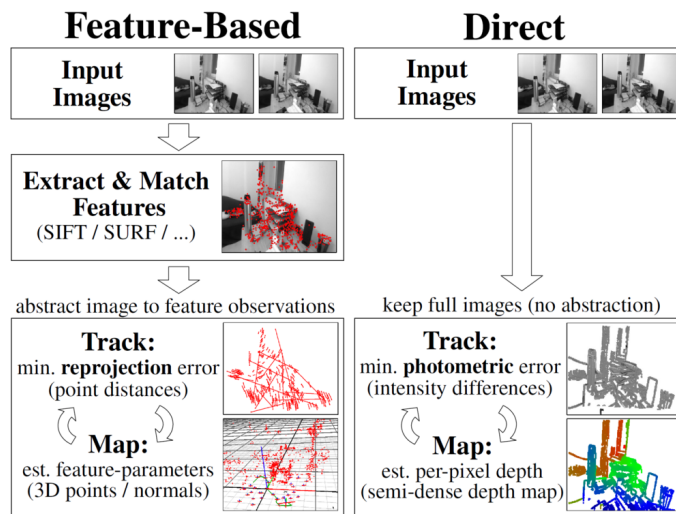
Feature-based and direct based methods have great advantages, a hybrid semi-direct method was proposed named semi-direct visual odometry (SVO) [130]. This method uses direct methods to obtain the pose, but also depends on characteristics of consistency. It uses a probability model for depth estimation and the deep filtering is based on a mixed model of Gauss and homogenous distribution.

In SVO, first the direct method is used to solve pose matching, then the Lucas-Kanade optical flow matching [131] to obtain subpixel accuracy and finally the reprojection error minimization is optimized by combining the point cloud map. The method relies on selecting key frames and does not directly match the whole image but instead extracts an image block to obtain the camera pose.

In conclusion, VO is considered a building block of visual SLAM. Although direct based methods are popular the key issues remain the lack of speed and consistency. Feature-based

and hybrid semi-direct based methods are able to build sparse maps and direct methods can build semi-dense maps. Figure 2.14 summarizes the two main approaches discussed, feature based versus direct methods. The feature point methods have been widely applied, although the description of feature points is mainly responsible for its accuracy. The direct methods are relatively recent technique that has strong robustness and may be applied to situations with very few features, such as hallways or smooth walls [132].

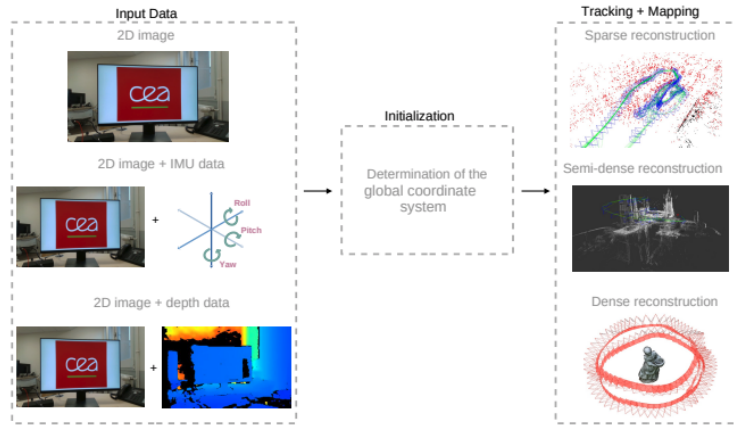
Figure 2.14 – Feature based vs Direct method pipelines [125].



2.3.2 Visual-SLAM

We now take a look at the next general concept that includes visual odometry algorithms, visual SLAM. Figure 2.15 shows the general visual SLAM pipeline where the data input ranges from 2D images, 2D images + Inertial Measurement Unit data or 2D image + depth data. The main three modules can be divided into input data, initialization, and tracking and mapping which are odometry algorithms that we looked at in the previous sections.

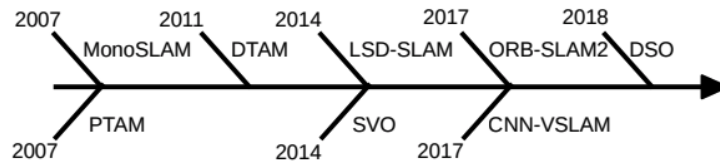
Figure 2.15 – Visual SLAM general pipeline [133].



Initialization is used to determine the global coordinate system to build an initial map which is then used to track and map. Tracking estimates the sensor’s pose continuously and establishes a correspondences between the frame and the map. This problem is called perspective-n-point, which is to estimate the pose of a calibrated camera from a set of n 3D points in the world coordinate system and 2D projected points in an image. Mapping is the process of computing and expanding the 3D environment, and can result in sparse, semi-dense, or dense 3D reconstruction.

Focusing only on the Visual (*only*) SLAM algorithms, figure 2.16 is a timeline of the most representative ones. We will briefly discuss each of these algorithms that mostly precede the incorporation of deep learning until 2017 [133, 134].

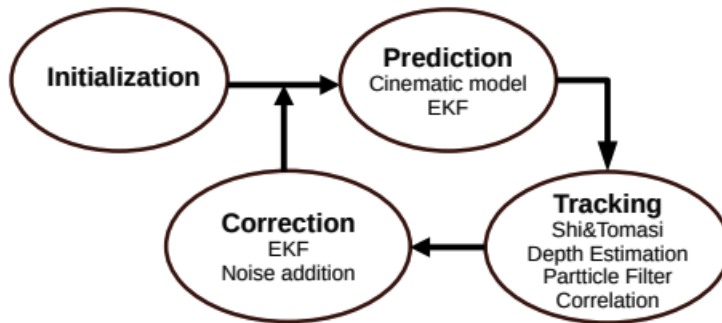
Figure 2.16 – Visual SLAM algorithms timeline [133].



MonoSLAM presented by David *et al.*[135]: this simple algorithm is a monocular SLAM algorithm where the first step is initialization. This is followed by prediction to update the

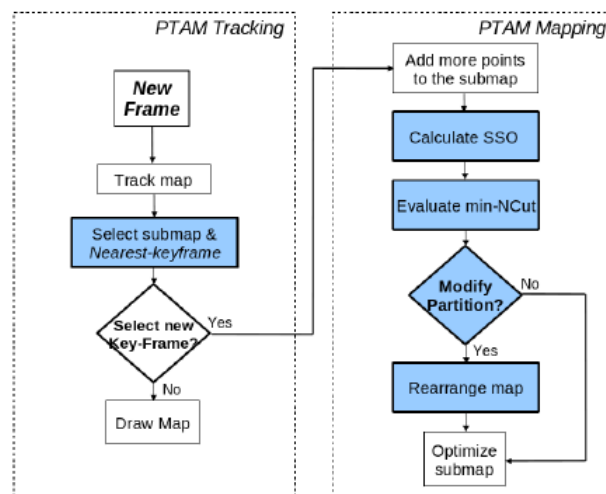
state vector by estimate the camera motion and environment structure using the extended Kalman filter (EKF). The algorithm estimates and updates in real-time and one constraint is the requirement of a known target for the initialization. The complexity depends on the size of the environment. Another constraint is that it only reconstructs a map of landmarks.

Figure 2.17 – MonoSLAM [133].



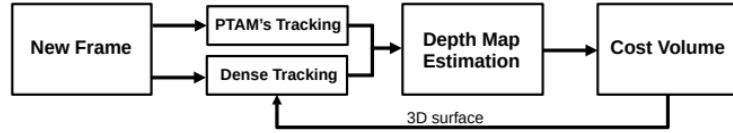
PTAM parallel Tracking and Mapping (PTAM) [136] is the first to separate the process of tracking from mapping. As shown in figure 2.18 the keyframes are added as the camera moves and the map that is initialized through the mapping process is expanded. PTAM computes camera poses by minimizing the reprojection error and the disadvantage of the algorithm is due to the bundle adjustment step required.

Figure 2.18 – PTAM SLAM [136].



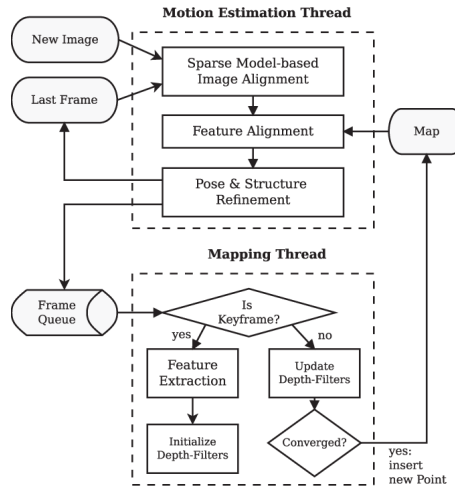
DTAM dense tracking and mapping proposed is a fully direct method [121]. There are two main parts as with any SLAM algorithm but in DTAM they are dense tracking, and dense mapping. The first steps results in depth map

Figure 2.19 – DTAM [121].



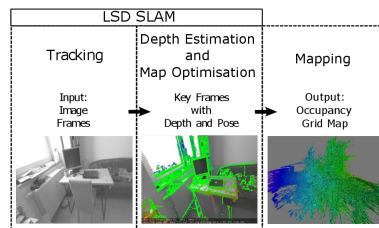
SVO the semi-direct visual odometry is a combination of feature and direct based methods. Motion estimation uses the photometric error and the mapping thread is based on probabilistic depth filters. The feature extraction is not needed for every frame and so the capable of operating with a high frame rate and low-cost embedded systems.

Figure 2.20 – SVO [130].



LSD-SLAM large-scale direct monocular SLAM is a direct algorithm that produces a semi-dense reconstruction. It has three main steps: tracking, depth map estimation and map optimization. It starts with photometric error to estimate the pose then keyframe selection. Finally the map is optimized through a pose-graph optimization algorithm.

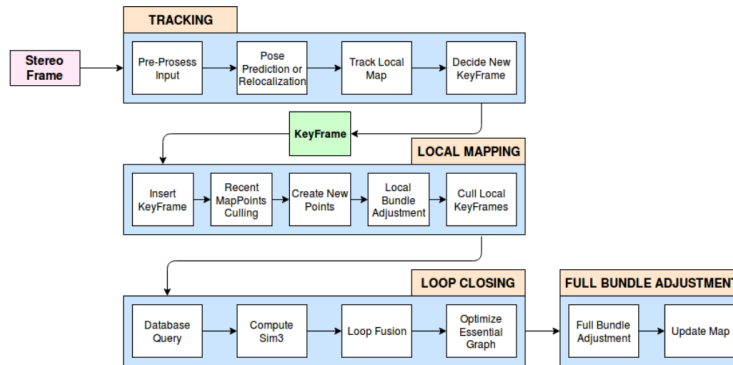
Figure 2.21 – LSD-SLAM [125].



ORB-SLAM2 consists of three main simultaneous threads: tracking, local mapping and

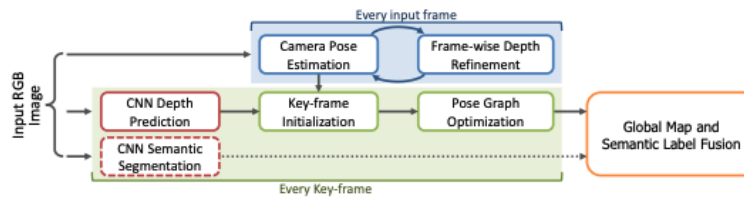
loop closing. Tracking finds the sensor’s pose by minimizing reprojection error. Local mapping manages map operations and lastly loop closing detects new loops and rectifies any drift error. After a full process is completed the algorithm considers the whole structure and uses bundle adjustment to estimate motion consistency.

Figure 2.22 – ORB-SLAM 2 [116].



CNN-SLAM is one of the earliest real-time SLAM systems based on convolutional neural networks. With two pipelines as shown in figure 2.23, there is a process for every input frame and another for every key frame. The top process predicts the camera pose by minimizing the photometric error between the current frame and the nearest keyframe. The bottom process predicts depth for every key frame, and the semantic segmentation. Pose-graph optimization is performed to obtain a globally optimized pose estimation. The CNN is trained with ground-truths of camera trajectory and depth maps.

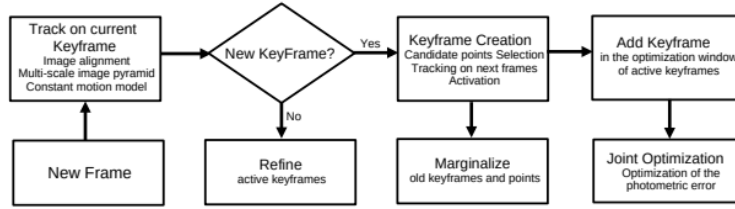
Figure 2.23 – CNN-SLAM [137].



DSO

Direct sparse odometry is a direct method that results in a sparse reconstruction. By applying local bundle adjustment that optimizes keyframes window and inverse depth map, the algorithm performs continuous optimization. The images are divided into blocks and only the highest intensity points are selected.

Figure 2.24 – DSO-SLAM [122].



The table below summarizes the visual-SLAM methods discussed briefly in this section. The nature of the method lines up with our objective of using deep learning to obtain a reconstruction of the environment. CNN-SLAM [137] is a supervised method but it is the ideal framework for us to follow and the first algorithm that incorporates convolutional neural networks to solve the SLAM problem.

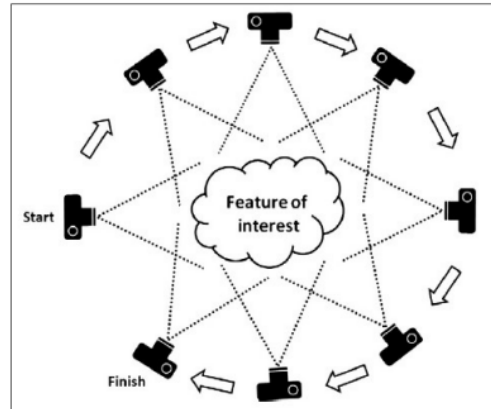
Table 2.1 – Visual-SLAM methods.

No.	Method	Type	Map	Reference
1	MonoSLAM	Feature-based	Sparse	[135]
2	PTAM	Feature-based	Sparse	[136]
3	DTAM	Direct	Dense	[121]
4	SVO	Hybrid	Sparse	[130]
5	LSD	Direct	Semi-dense	[125]
6	ORB-SLAM 2	Feature-based	Sparse	[116]
7	CNN-SLAM	Direct	Semi-dense	[137]
8	DSO	Direct	Sparse	[122]

2.3.3 Structure from Motion

The field of structure from motion (SfM) which is a more general concept encloses visual-SLAM and visual odometry. SfM is the notion of creating a map using several images taken from different perspectives, or even different cameras. The 3D structure of the environment can be determined from a set of multiple overlapping images from a moving camera as shown in figure 2.25. The principle of SfM remains that the location and pose of the camera (s) must be known. In some cases the absence of GPS or sensors that give such information, triangulation is used to reconstruct scene geometry. This brings us back to the the requirement of projective geometry which is the basis of solving this problem. Some of the steps in an SfM offline algorithm often mimic the building blocks of a fundamental SLAM algorithm.

Figure 2.25 – SfM [138].



Early self-calibrating metric reconstruction systems [139–142] are the foundation for systems that followed. A technique to address the SfM issue without the requirement for a priori correspondence knowledge was described by Dellaert et al. in [143]. It can handle photographs presented in any sequence and shot from a variety of different angles. Finding the structure and motion with the highest likelihood given just the 2D measurements entails combining all potential 3D feature mappings to 2D measurements. An technique that iteratively refines a probability distribution across the set of all correspondence assignments was used to get the desired result. The resulting method is quick, straightforward, and easy to use. An approach to estimate structure and motion using a series of photos gathered in a causal manner was given by Jin et al [127]. A class of geometric and photometric models for the scene that may be finitely parameterized is used by the algorithm to integrate visual input. A closed loop is created by monitoring the picture region and combining 3D motion estimates. They framed the SfM issue within the context of nonlinear filtering. By recreating the state of a nonlinear dynamical system using an extended Kalman filter, the unknown structure and velocity are estimated. Furthermore, they have demonstrated that the dynamical system is observable when the translational velocity is non-zero, the scene comprises at least two planar patches, each with a distinct normal direction. The algorithm’s recursive structure makes real-time implementation possible.

Agarwal et al. [144] developed a method to convert massive, disorganized collections of pictures into 3D geometry. The system uses image matching and 3D reconstruction methods that maximize parallelism at each pipeline level. The system is broken down into two parts: (1) Pre-processing, where photos are stored centrally and then delivered to cluster nodes as needed in chunks of a specific size. Each node extracts SIFT features while downsampling its images to a set size. (2) Validation: suggest possible image pairs and validate them (using feature matching) (3) Track generation: aggregate these features so that a single 3D point may

be estimated from all of them by the geometry estimation algorithm. A group of computers (called nodes) make up the system, and one of them is designated as the master node, which manages work scheduling. A strategy for unstructured image collections that takes into account every photo at once rather than developing a solution piece by piece was provided by Crandall [145]. The method computes an initial estimate of the camera position from all available images and then uses bundle adjustment to improve that estimate for scene structure. The method employs a two-step procedure. Levenberg-Marquardt non-linear optimization, which is related to bundling adjustment but involves extra constraints, is used to estimate camera parameters in the second phase after the discrete belief propagation (BP) technique in the first step. When compared to current incremental bundle adjustment (IBA) techniques, the method provides superior reconstructions and is faster. Numerous SfM techniques have been proposed, including global [145–147], hierarchical [148], and incremental [144, 149–151].

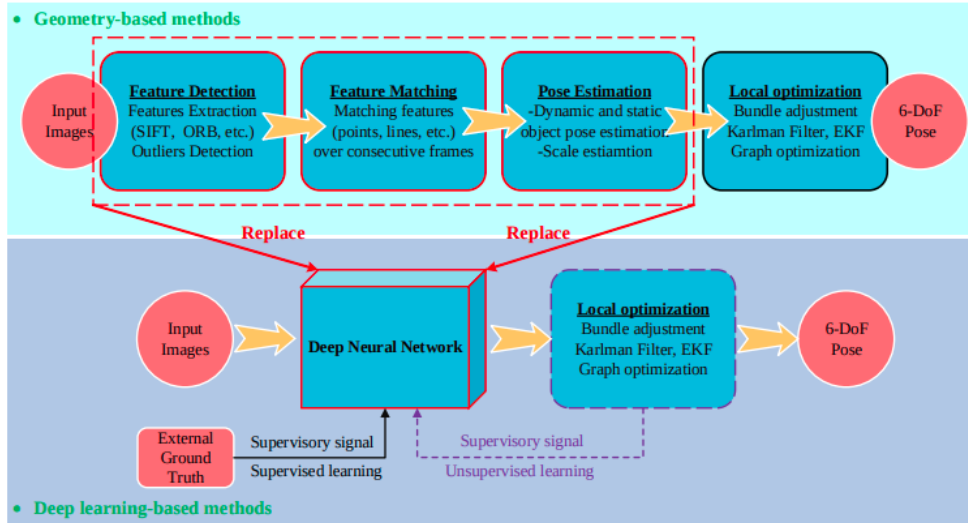
2.3.4 Conclusion

More detailed surveys on all three fields, and interesting literature that details these offline algorithms can be found in [152–159]. These methods are not only more time-consuming, but also more complex to implement in order to fix our issue. Additionally, they rely significantly on offline knowledge and standardized equipment. Even in this field, there is a clear trend toward incorporating deep learning to improve the accuracy and automation of camera motion estimations. The tendency is also supported by enhanced performance, reproducibility, and accessibility. Therefore, as mentioned in the objectives of this thesis, we want to solve our VO problem with the deep learning. Therefore we focused our attention on visual odometry with deep learning looking at proposed methods in detail, their applications and what is prominent in this division of the field.

2.4 Visual Odometry with Deep Learning

As we have seen the classical geometry-based VO field is impressive and has progressed over the years. The robustness of these methods continue to be a challenge in certain environments, and their heavy computational cost is a persisting disadvantage. Many researchers aim to use deep learning techniques to the VO problem in an effort to lower the high computing cost; their work can be separated into supervised and unsupervised methods. An example of what deep learning can replace is shown in Figure 2.26, ultimately estimating the camera pose from data directly. As mentioned in Section 2.1.1 ground truth is required as a supervision signal for supervised methods while the output is used as supervision signal in unsupervised methods.

Figure 2.26 – Geometry-based VO & Deep learning-based VO [160].



Deep learning solutions have been employed for feature detection, matching, pose estimation and depth estimation. Feature extraction and matching methods require careful design and special fine-tuning to work well in certain environments. Given the quality of our data and the difficulty to not only extract meaningful features but also match them we focused on methods that estimate pose in other ways. We also refined our search to unsupervised methods, including the subsection of self-supervised methods. By estimating depth and pose the goal is to mimic how we humans are able to perceive new scenes from a single monocular image. We first review datasets and evaluation indicators used in depth estimation for monocular depth and camera pose or egomotion estimation [161].

KITTI dataset [162, 163], is the most common for computer vision sub-tasks such optical flow [164], visual odometry [165], depth [166], object identification [167], semantic segmentation [168], and tracking [169]. It is also a frequent benchmark and training dataset for unsupervised and semi-supervised methods. The 56 scenes presented in the dataset are separated into 28 scenes for training and the remaining 28 scenes for testing by Eigen et al. [166]. The dataset contains real images from “city”, “residential” and “road” categories. Stereo picture pairs in each scene have a resolution of 1224 x 368. A revolving LIDAR sensor sparsely samples the corresponding depth of each RGB frames. This provides ground truth for methods trained in a supervised manner. It also offers the posture ground truth for 11 odometry sequences, it is frequently used to assess deep learning-based visual odometry (VO) algorithms [170, 171].

Cityscapes dataset [172] is mainly used for semantic segmentation tasks [172]. In this dataset, there are 20,000 photos with coarse annotations and 5,000 images with fine annotations. It comprises of a collection of stereo video sequences that have been gathered over many months

Figure 2.27 – Kitti dataset samples [162].



from 50 cities.

The dataset is used for training of numerous unsupervised depth estimation algorithms [173, 174] because there is no depth groundtruth, but disparity is provided which can allow for depth estimation. Pre-training depth networks on Cityscapes enhances their performance [173–176]. The dataset is comprised of 22,973 stereo picture pairings with a resolution of 1024 x 2048.

Figure 2.28 – Cityscape dataset samples [177].



A widely used evaluation method with five evaluation indicators—RMSE, RMSE log, Abs Rel, Sq Rel, and Accuracies—is provided in [166] in order to assess and compare the performance of various depth estimation networks. These indications are as follows:

- **RMSE** = $\sqrt{\frac{1}{|N|} \sum_{i \in N} \|d_i - d_i^*\|^2}$,
- **RMSE log** = $\sqrt{\frac{1}{|N|} \sum_{i \in N} \|\log(d_i) - \log(d_i^*)\|^2}$,
- **Abs Rel** = $\frac{1}{|N|} \sum_{i \in N} \frac{|d_i - d_i^*|}{d_i^*}$,
- **Sq Rel** = $\frac{1}{|N|} \sum_{i \in N} \frac{\|d_i - d_i^*\|^2}{d_i^*}$,
- **Accuracies**: % of d_i s.t. $\max(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}) = \delta < thr$,

where d_i is the predicted depth value of pixel i , and d_i^* stands for the ground truth of depth. N is the total number of pixels with real-depth values, and thr is the threshold.

2.4.1 Unsupervised deep learning-based methods

The geometric restrictions between frames are used as the supervisory signal during training of the unsupervised methods in place of the ground truth. Following a stereo camera baseline, the overlapped area between two stereo images, each pixel in one image can find its correspondence in the other with the horizontal distance H [157].

$$H = Bf/D, \quad (2.9)$$

where B is the baseline of a stereo camera, f is the focal length, and D is the depth value of the corresponding pixel. This can be translated to monocular systems. A fundamental unsupervised model is as follows: the geometric constraints for unsupervised algorithms is based on the projection between adjacent frames, and they are learned from a monocular image sequences:

$$p_{n-1} \sim KT_{n \rightarrow n-1}D_n(p_n)K^{-1}p_n, \quad (2.10)$$

p_n stands for the pixel on image I_n , and p_{n-1} to the corresponding pixel of p_n on image I_{n-1} . K is the camera intrinsics matrix, $D_n(p_n)$ the depth value at pixel p_n , $T_{n \rightarrow n-1}$ represents the spatial transformation between the two images I_n and I_{n-1} .

Figure 2.29 – Warping process for view reconstruction in unsupervised methods [174–176].

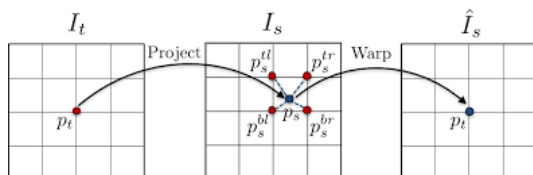


Figure 2.29 illustrates the differentiable image warping process where for each point p_t in the 'target' view, it is projected onto the 'source' view based on predicted depth and camera pose. Bilinear interpolation is then used to obtain the value of a new warped image \hat{I}_s .

Zhou *et al.*[176] estimates $D_n(p_n)$ and $T_{n \rightarrow n-1}$ using a depth network and a pose network. The networks predict the depth map \hat{D}_n from a single image I_n , and a pose network to regress the transformation $\hat{T}_{n \rightarrow n-1}$ between frames (I_n and I_{n-1}). Pixels correspondences are established by the projection function and based on estimations, the correspondences between I_n and I_{n-1} are developed:

$$p_{n-1} \sim K\hat{T}_{n \rightarrow n-1}\hat{D}_n(p_n)K^{-1}p_n. \quad (2.11)$$

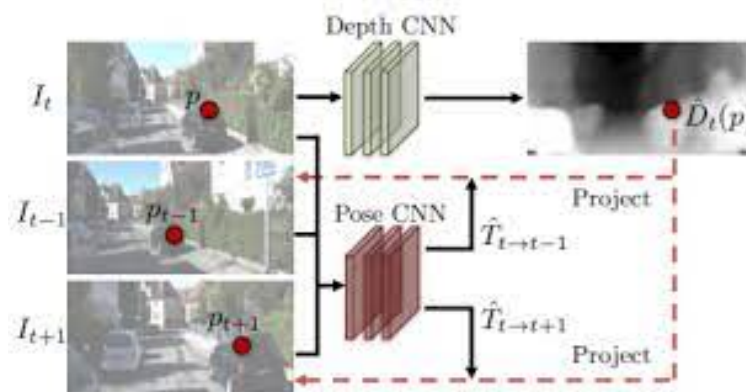
The supervision signal which serves as the geometrical constraint is calculated as the photometric error between corresponding pixels. Inspired by [178] the reconstruction loss is formulated

as :

$$\mathcal{L}_{vs} = \frac{1}{N} \sum_p^N |I_n(p) - \hat{I}_n(p)|, \tag{2.12}$$

where p indexes over pixel coordinates and $\hat{I}_n(p)$ the reconstructed frame.

Figure 2.30 – Unsupervised learning of depth and egomotion pipeline [176].



2.5 Conclusion

In this chapter we first presented the three main concepts we will exploit in this work. We also presented existing deep learning-based methods for DED quantification. There is a contrast between the existing methods and the concepts we wish to utilise which have yet to be deployed for DED diagnosis. We detail the existing methods for Odometry, and Visual Odometry which is a very rich field. Lastly, we present the most recent contributions in the field with deep learning. We observed a general tendency toward the use of deep learning-based, and we concentrate on unsupervised methods since we believe they are especially suitable for our work. Therefore we will learn to predict the camera movements from our data without the use of any sensors. We continue on to detail the materials in the following chapter to look at the resources we are using for this thesis.

MATERIALS

“Without data, you’re just another person with an opinion.”

— *W. Edwards Deming*

3.1	Introduction	56
3.2	Acquisition Method	56
	3.2.1 Original Database 'O'	56
	3.2.2 PEPSS Database 'P'	57
3.3	Camera Calibration	59
3.4	Dataset	60
3.5	Ground truth annotations : PEPSS Database 'P'	61
3.6	Conclusion	64

This chapter details the materials collected and used for this thesis. We look at the different protocols and the acquisition methods used to obtain the database. Calibration of the camera was also necessary for our framework. We discuss the methods and results of camera calibration and conclude with the finalised dataset.

3.1 Introduction

When LaTIM started studying the automation of DED diagnosis (Master 's thesis funded by Laboratoires Théa in early 2019), a preliminary protocol was set up and a dataset of multi-factorial DED patients was collected. This initial dataset is referred to as Original database 'O'. In the framework of Necessity¹, there was a prospective cohort PEPSS (Recueil des symptômes et évaluation de la sécheresse oculaire : développement de nouveaux outils pour le syndrome de Sjögren primitif) that evaluated the ocular surface damages in patients with Sjogren's syndrome.

This was a slightly improved protocol where we also collected a dataset of DED examinations from Sjogren's syndrome patients. This second dataset is referred to as PEPSS database 'P'. Both datasets have similarities, and so we were able to use them jointly in some experiments All examinations were conducted at the Service d'Ophthalmologie, Brest University Hospital Centre (CHRU Brest) by Ophthalmologists Dr. Anas-Alexis Benyoussef and Dr. Beatrice Cochner. We will go into further details on how we used each of these for our experiments and proposed method in Chapter 5.

3.2 Acquisition Method

The videos were recorded using the Haag Streit BQ 900 slit lamp² and the camera module CM 900 (*resolution* : 1600×1200 pixels, 12 frames/second). A Galilean microscope with a magnification range of 6.3 to 40 that is adjustable in 5 fixed steps is included as standard equipment with the BQ 900 as shown in Fig.3.3a. The slit lamp has an improved clinical vision through the light transmission and the optical quality. It also includes the imaging systems : IM 600 and IM 910 cameras. The EyeSuite software Fig.3.3b controls all Haag-Streit devices and is made to optimize patient care in busy practices. It allows for access to patient data, patient management system and more importantly for our objective image and video capture. Following the Haag-Streit image exposure guide three modules, shown in Fig.3.4, were used.

3.2.1 Original Database 'O'

The collection of the original database was performed at the CHRU Brest service. The dataset contains two magnification settings: x10, x16. During these examination multiple settings were tested until a satisfactory acquisition was obtained that allowed a complete DED grading. This was done by Dr. Benyoussef, Mathieu Lamard, and Bendy Latortue who worked giving a

1. <https://www.necessity-h2020.eu/>

2. <https://www.haag-streit.com/haag-streit-diagnostics/products/slit-lamps/bq-900/>

clinician's and an engineer's input. Patients that were examined had DED symptoms and so the protocol was the following :

1. Blue light yellow filtered fluorescein sodium staining :
 - (a) Three blinks with a pause to view the tear film breaking (TBUT diagnosis)
 - (b) Cornea grading
 - (c) Temporal and nasal conjunctiva grading.
2. White light lissamine green staining:
 - (a) Temporal and nasal conjunctiva grading.


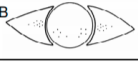



3.2.2 PEPSS Database 'P'

For the examinations under the PEPSS study was set to x 10. This study included patients that had been diagnosed with Sjogren's syndrome. The study also included multiple diagnostic examinations specific to Sjogren's syndrome. Given that dry eye is also a principal symptom the ocular surface was analysed and the damage was graded through staining examinations. This included illumination with white light (lissamine green evaluation) followed by cobalt blue light and interposition of a yellow filter (fluorescein evaluation). There were two clinical diagnostic tests for DED, tear break up time measurement and superficial punctate keratitis (SPK) grading that can be done using various grading scales as mentioned in Chapter 1. The protocol followed was the following:




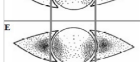
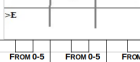
1. Schirmer's test
2. Blue light yellow filtered fluorescein sodium staining:
 - (a) Three blinks with a pause to view the tear film breaking (TBUT diagnosis)
 - (b) Cornea grading following the Oxford scale.
 - (c) Temporal and nasal conjunctiva grading following the Oxford scale.
3. White light lissamine green staining:
 - (a) Temporal and nasal conjunctiva grading following the Oxford scale.
4. Patient questionnaire Ocular Surface Disease Index (OSDI).

The staining grade was given using the modified Oxford scale for both cornea and conjunctiva, shown in Figure 3.2. OSDI is a 12-item questionnaire that assesses dry eye symptoms where patients rate their responses on a 0 to 4 scale and the final score ranges from 0 to 100.

Figure 3.1 – Oxford scale [5].

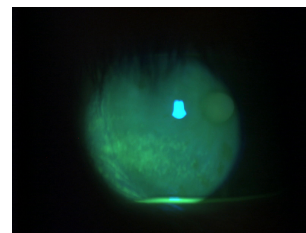
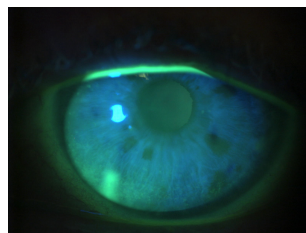
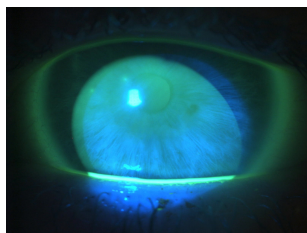
PANEL	GRADE	CRITERIA
A 	0	Equal to or less than panel A
B 	I	Equal to or less than panel B, greater than A
C 	II	Equal to or less than panel C, greater than B
D 	III	Equal to or less than panel D, greater than C
E 	IV	Equal to or less than panel E, greater than D
>E	V	Greater than panel E

(a) The Oxford grading scale [5].

PANEL	Grade	Criteria
A 	0	Equal to or less than panel A
B 	I	Equal to or less than panel B, greater than A
C 	II	Equal to or less than panel C, greater than B
D 	III	Equal to or less than panel D, greater than C
E 	IV	Equal to or less than panel E, greater than D
>E	V	Greater than E
FROM 0-5	FROM 0-5	FROM 0-5
TOTAL SCORE:		FROM 0-15

(b) Modified Oxford scale [5].

Figure 3.2 – Examples of DED grading.

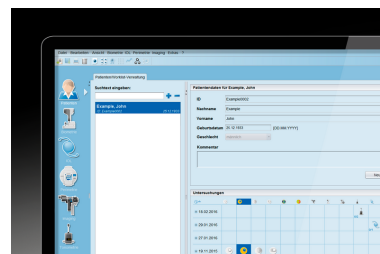


(a) Cornea Oxford grade = 1. (b) Cornea Oxford grade = 3. (c) Cornea Oxford grade = 4.

Figure 3.3 – Acquisition method.

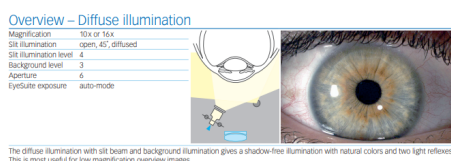


(a) Haag Streit BQ 900 slit lamp

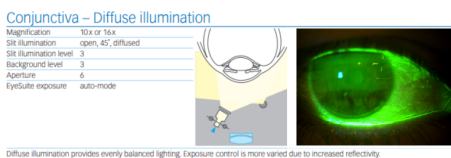


(b) Eyesuite Software

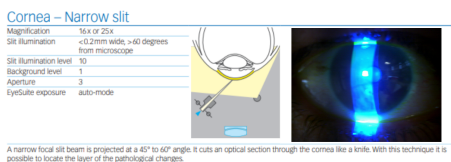
Figure 3.4 – Haag-Streit image exposure guide.



(a) Overview – Diffuse illumination



(b) Conjunctiva – Diffuse illumination



(c) Cornea – Narrow slit

3.3 Camera Calibration

We performed camera calibration using two different checkerboards, 8 x 7 squares and 9 x 10 squares both 1mm. Figure 3.2 and 3.3 show both checkerboards photos and with a better visualisation of the 8 x 7 checkerboard we continued with the calibration using those set of images. The images of the 9 x 10 checkerboard often had areas that were covered which is mainly due to the size of the checkerboard and the small field of view of the camera. Given the protocol used we conducted all the camera calibration image acquisitions at a magnification of x10. We conducted two experiments to obtain the camera intrinsic parameters using MATLAB R2020a and a Python code using OpenCV. Using the Camera Calibrator application and frames of the filmed checkerboard. The videos filmed resulted in a number of extracted frames ranging from 1080 to 1800, depending on the length of the video. Both methods employed include frame selection, where only around 5% of the frames extracted were retained. This can mainly be attributed to the quality of the image aswell as the chessboard print being flat and non-spherical like the eye. The majority of the rejected frames had blurry parts of the chessboard, making it impossible to detect the corners. The experiments were also done using the raw image size obtained from the video camera (1600x1200 pixels) and resized images (192x256 pixels). The resized images were what we used for all the baseline experiments and our proposed method training which will be detailed in chapter 5.

In order to compare the two methods; MATLAB & OpenCV, we calculated the re-projection

Figure 3.5 – 9 x 10 checkerboard photos through Haag Streit BQ 900 slit lamp.

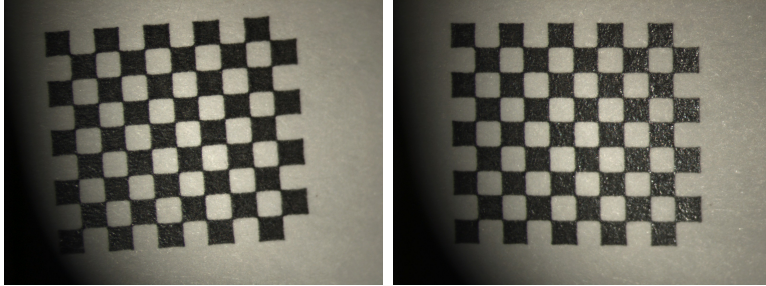
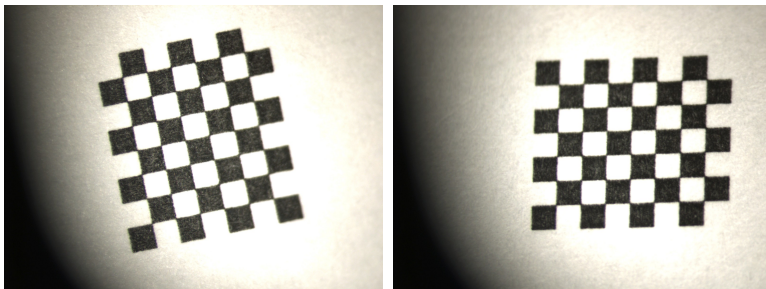


Figure 3.6 – 8 x 7 checkerboard photos through Haag Streit BQ 900 slit lamp.



error which is commonly used to evaluate the camera calibration of a single camera. Camera calibration is done by measuring feature points with a known spatial relation to each other. Assuming a pinhole camera model, we use the checkerboard size to detect the points, origin and ultimately the re-projected points. Once the feature points are detected we measure the spatial inter-relation and the more valid images we have the better our measurements of the intrinsic and extrinsic parameters are. We then re-project the feature points onto the scene using the camera model and compute the re-projection error as an average L2 norm of all point correspondences :

$$\text{Reprojection error} = \frac{1}{N} \sum_{i=0}^{N-1} |p_i - q_i|_2 \quad (3.1)$$

where p_i are the observed points and q_i are the feature points locations predicted on the image plane. For both image sizes the lowest re-projection errors we obtained were using the Python code calibration implementation, although values were fairly close with the Matlab toolbox.

3.4 Dataset

The original database of videos, we labelled 'O' contains 79 videos of unique eyes. Although they were of poor quality and did not follow the same protocol as those acquired for the PEPSS study, they were still useful for various tasks that will be detailed in chapter 5. The lack in

Figure 3.7 – Reprojection Error

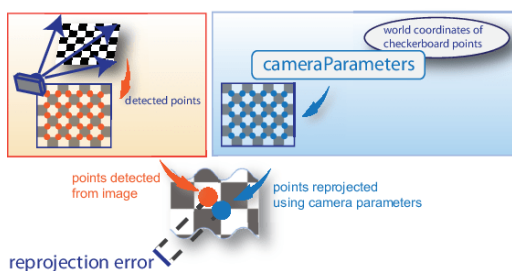


Table 3.1 – Camera calibration results

Method	Image Size	Fx	Fy	Cx	Cy	Reprojection Error
Python	1600x1200	29,856.2	29,367.7	800.0	600.5	0.27
Matlab	1600x1200	33,172.8	32,981.5	622.7	539.9	1.37
Python	192x256	3,758.9	3,758.9	138.8	85.4	0.07
Matlab	192x256	5,919.2	5,793.8	76.8	89.5	0.17

quality present in the original database includes changes in lighting fixtures, zoom parameters and also more abrupt motions which were all necessary to finalise the protocol. Database 'O' was also used for an internship in 2019 with similar scope and achieved good results for classification of open/closed eye. As part of the PEPSS study 39 patients were evaluated but we only obtained 26 examinations. We named this database 'P' and it contained 52 videos of unique eyes. Table 3.2 summarises the two databases and shows examples of frames taken from a few examinations of each. Figure 3.8 shows the two box plots with the minimum, first quartile, median, third quartile, and maximum of both databases for the time and number of frames.

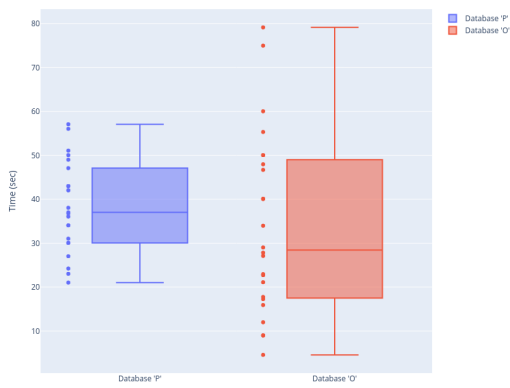
3.5 Ground truth annotations : PEPSS Database 'P'

Considering the DED specific diagnostic tests conducted under the PEPSS study, we obtained various ground truth annotations for all the data. Figure 3.9 shows a normalized box-plot of these grades. We wanted to focus on using the oxford score for our application where the data seemed to be variable.

For more precise annotations we asked five experts and Dr. Benyoussef to manually annotate the punctate dots. The damaged areas that appear in fluorescent green is what the aims to oxford score quantify. In order to render this more precise we obtained precise co-ordinate locations

Figure 3.8 – Databases box-plots

(a) Databases vs. Video Length(sec)



(b) Databases vs. No. of frames

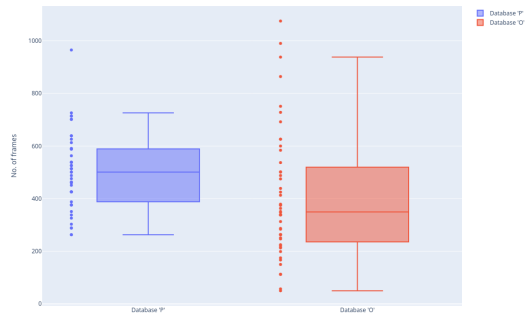


Figure 3.9 – Diagnostic test results description.

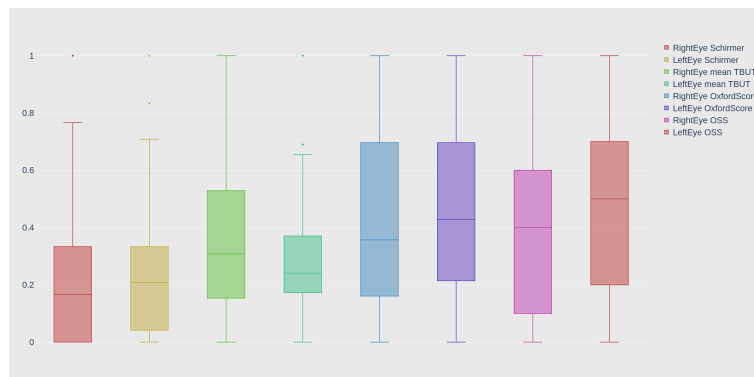
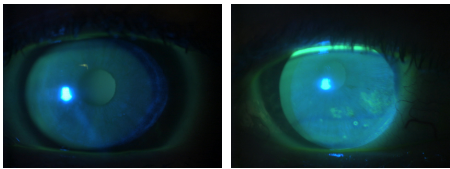
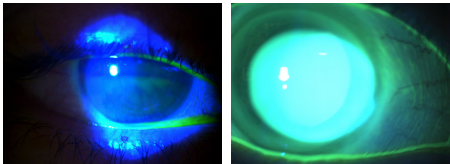
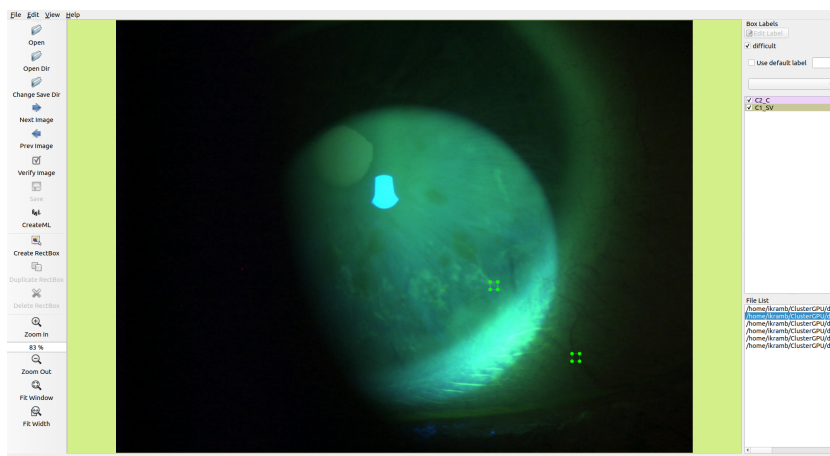


Table 3.2 – Details of Databases

Database	No. Videos	Total frames	Examples
PEPSS Database 'P'	52	23,023	
Original Database 'O'	47	19,532	

of the punctate dots that were visible on the surface of the eye. For a set of three patients we randomly chose 66 frames to annotate, treating them as pairs with a random step of n in between. Our final set consists of 126 points on the sclera on vein intersections and punctate dots, and 33 points on the cornea of visible punctate dots. This was conducted using the graphical image annotation tool LabelImg [179]. The tool creates bounding boxes and we take the middle point as the co-ordinates. The figure 3.10 is an example of the procedure described showing two kinds of annotation; cornea punctate dot and a sclera vein intersection. The ground truth is then saved for each patient in an easily accessible JSON file format.

Figure 3.10 – LabelImg tool [179].



After the first annotations of the areas on two frames, the following experts were given the annotations of the first frame and asked to annotate the second frame by finding the same areas. This allowed us to calculate what we considered a human error. By taking the primary co-

ordinates as ground truth, we compared all the rest to it and calculated the difference between all the graders through the euclidean distance Eq.3.2. The results are shown below :

$$euclidean\ distance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (3.2)$$

where x_1, y_1 are the ground truth points, and x_2, y_2 are the secondary grader's points we want to compare.

Table 3.3 – Grader error results.

Grader errors	Mean Euclidian (px)	Mean Euclidian (%)
All annotations	5.30	0.33
Annotations on sclera	5.68	0.35
Annotations on cornea	3.33	0.21

3.6 Conclusion

In this chapter we detailed the two databases we have and the camera calibration we implemented. Given that both databases use the same camera, the intrinsic parameters we obtained could be applied when using either databases. Following this we go on to a few main baselines and analyse their main contributions and how they correlate to their objectives.

BASELINE METHODS

“Research means that you don’t know, but are willing to find out”

— *Charles F. Kettering*

4.1	Introduction	66
4.1.1	Depth prediction for ego-motion	67
4.1.2	Joint depth & egomotion prediction	67
4.2	Conclusion	72

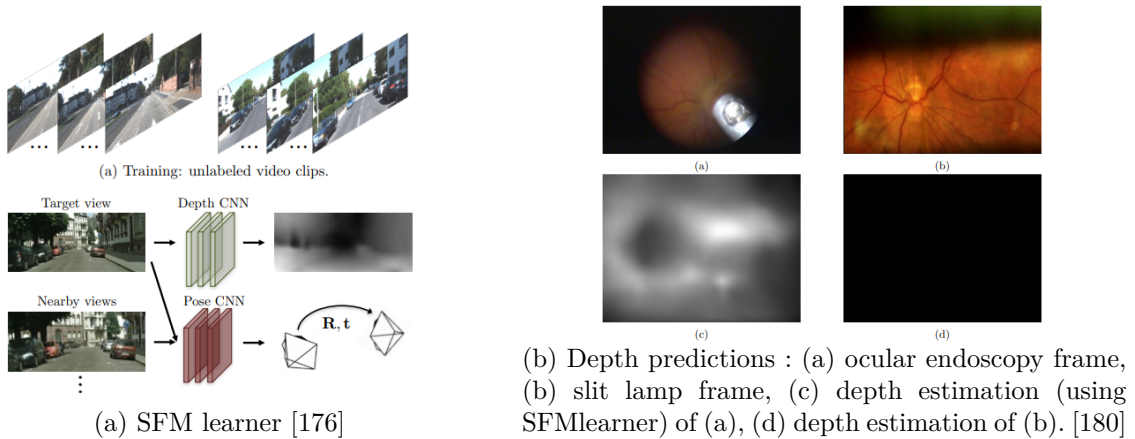
This chapter looks at the key baseline methods that we wanted to focus on. We take a closer look at the state-of-the-art and the techniques they implemented. Naturally, we carefully examine any assumptions made and anything that does not line up with our research topic.

4.1 Introduction

The extensive work done by Alexandre Guerre [180] was the first to address augmentation of field of view for a ophthalmology problematic. The thesis work looked at both ocular endoscopy and, retinal videos but have the same light and motion disturbances as our data. Alexandre showed that CNNs such as FlowNet were not very promising [181], given that the estimation of motion was very difficult and required ground truth data. A similar conclusion was reached by [182] given the light variations present in the dataset. Ultimately successful methods all required the creation of an artificial dataset with the ground truth for optical flow estimation. In our approach we wanted to implement a self-supervised or unsupervised method and avoid the creation of an artificial dataset.

A simplified SFM learner [176] Figure 4.1a was more encouraging, but the depth estimation was also difficult with both datasets Alexandre experimented with, results in Figure 4.1b.

Figure 4.1 – Depth estimations from Alexandre Guerre’s thesis [180]



As mentioned in Chapter 2 autonomous driving is a well known topic to benefit from AI potential. Just as SFM learner used the Kitti database, there are methods that followed by Zhou *et al.*[176] and Gordon *et al.*[183] with multiple improvements. Given the difficulty of estimating depth alone we decided to focus on such methods that jointly learn depth and egomotion. The main supervision signal and equation we wanted to solve was the following :

$$p_s \sim K\hat{T}_{t \rightarrow s}\hat{D}_t(p_t)K^{-1}p_t \quad (4.1)$$

where p_s are pixels in the source frame, p_t are pixels in the target frame, K is the intrinsic matrix, and is the relative depth $\hat{D}_t(p_t)$.

Using this warping operation and given two frames ; source frame (s) and a target frame

(t), we can translate the scene to the next frame and obtain the next image by projection. In order to solve for p_s , we need to solve for $\hat{D}_t(p_t)$ and $\hat{T}_{t \rightarrow s}$. This can be accomplished through two steps :

1. Depth Estimation

$$D_i = \theta(I_i)$$

$$\theta : \mathbb{R}(H \times W \times 3) \rightarrow \mathbb{R}(H \times W)$$

2. Pose Estimation :

$$E_{1 \rightarrow 2} = \psi_E(I_1, I_2) = (t_x, t_y, t_z, r_x, r_y, r_z)$$

$$E_{2 \rightarrow 3} = \psi_E(I_2, I_3) = (t_x, t_y, t_z, r_x, r_y, r_z)$$

where H = height and W = width of the image. Where a dense depth map is estimated from a single RGB frame, and the pose estimation takes in a sequence of two RGB images as input and produces the transformation between the frames and giving both the translation (t_x, t_y, t_z) and rotation parameters (r_x, r_y, r_z) between the frames.

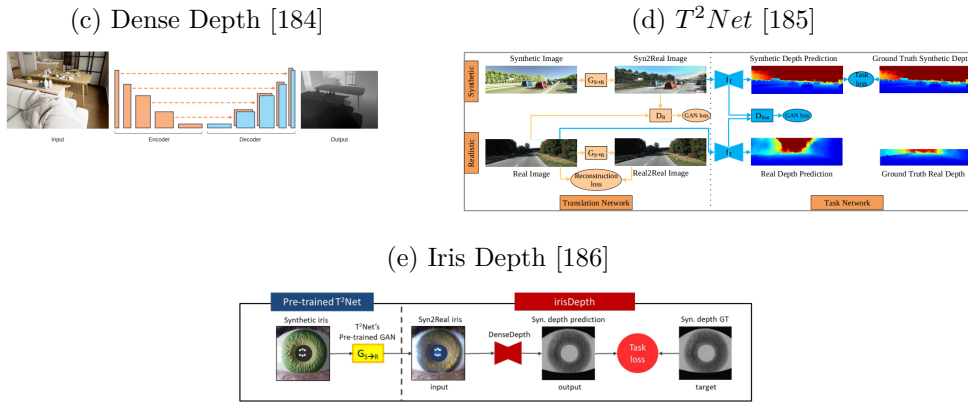
4.1.1 Depth prediction for ego-motion

We looked at existing methods for image depth estimation, although most require real image-depth or stereo images for training. A simple implementation is DenseDepth [184], with an encoder-decoder architecture. This served as a baseline for our depth map prediction task. Another proposed method T^2Net [185] is trained on synthetic image depth pairs and unpaired real images. It consists of a translation part and a task prediction part. Lastly, the method 'IrisDepth' closest to our problematic was proposed by Benalcazar *et al.*[186]. The paper focuses on obtaining a 3D iris scanner from a single image using CNN. Any 3D reconstruction requires depth estimation and this is the task we wanted to experiment with our data. IrisDepth is very similar to T^2Net as can be seen in their framework in Figure 4.2. We implemented each of the this methods, and T^2Net with only the task network training to predict the depth. All the depth map estimations are displayed in Table 4.1. Based on difficult training with losses diverging to infinity or not converging, and depth map results none of the methods seemed to work for our problematic.

4.1.2 Joint depth & egomotion prediction

Baseline Methods Depth learning can also be joined with egomotion, and this is also a rich field that we took a closer look at. We started with the a simple overview of SFM learner which is a framework for unsupervised learning used to estimate monocular depth and camera motion

Figure 4.2 – Depth estimation methods



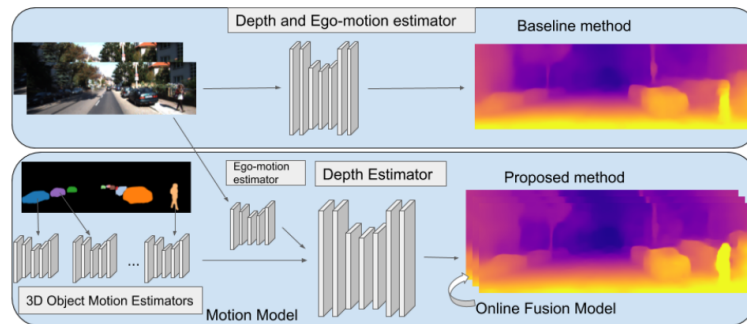
from unstructured video sequences. We included the two more recent methods by Gordon *et al.*[183] and Zhou *et al.*[176], that both learn motion transformations by enforcing photometric consistency. Lastly, the closest and most recent method with a medical application is Endo-SfMLearner [187].

Struct2Depth [176] tackles the challenging task of outdoor robot navigation and the proposed novel approach that models moving objects and introduces geometric structure by modeling the scene and the individual objects, shown in Figure 4.3a. The 3D object motion estimation is solved by a model with similar architecture as the ego-motion [188].

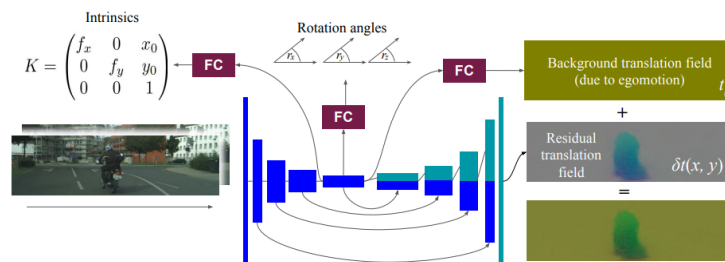
Depth from videos in the wild [183] establishes a new state of the art results on both depth and odometry predictions. The method also showed that depth can be learned from a collection of YouTube videos [183], shown in Figure 4.3b.

Endo-SfMLearner [187] first provided a new comprehensive endoscopic SLAM dataset. The synthetically generated ex-vivo dataset included six porcine organs and four acquisition methods. With both depth and pose annotations. The unsupervised method is able to obtain decrease in rotational errors and comparable results to SFM Learner, SC-SfMLearner, and Monodepth2 [175, 176, 189]. The method also addressed the similar disturbances of fast frame-to-frame illumination changes present in endoscopic videos through a brightness-aware photometric loss, shown in Figure 4.3c. The results for the depth map predictions are shown below in Table 4.2. Qualitative evaluation of the depth maps showed us very minimal improvement. Depth maps predicted with Endo-SfMLearner were able to capture more information than the other baseline methods.

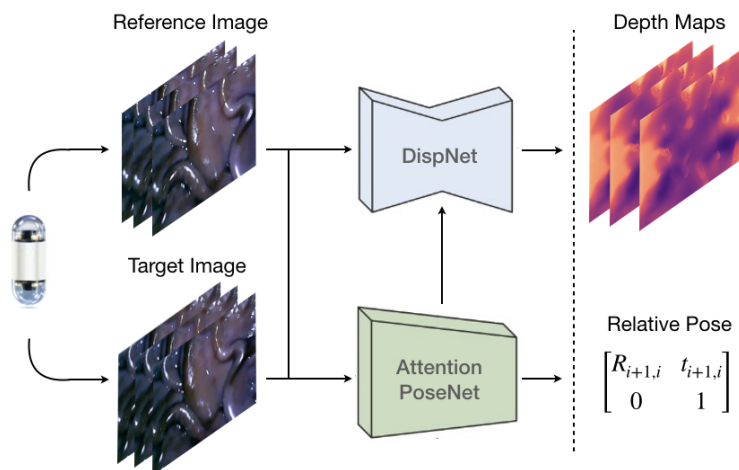
Figure 4.3 – Deep learning SLAM methods.



(a) Struct2Depth [188]



(b) Depth from videos in the wild [183]



(c) Endo-SfMLearner [187]

Table 4.1 – Depth map predictions part I.

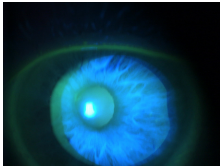
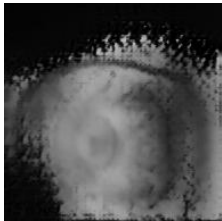

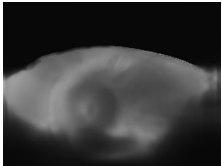
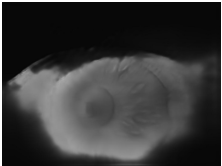
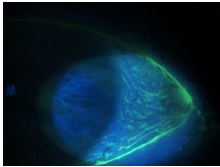
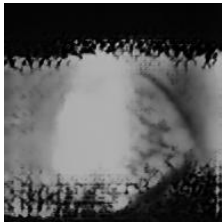
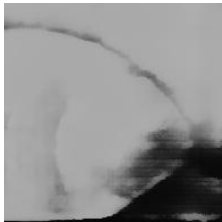
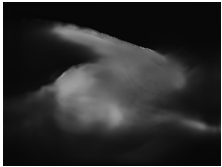
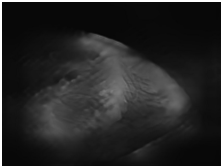
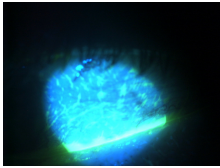
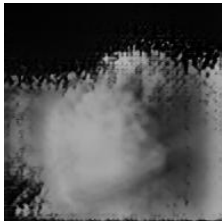


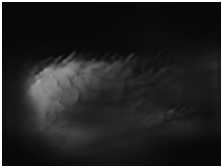
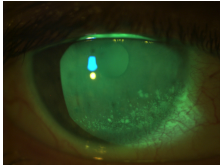
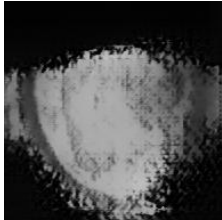
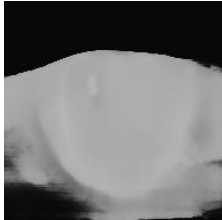
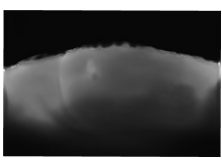
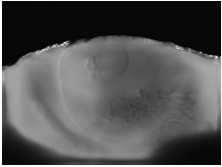
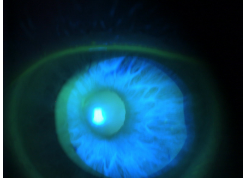
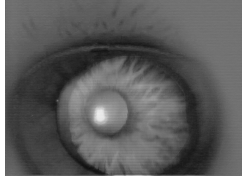
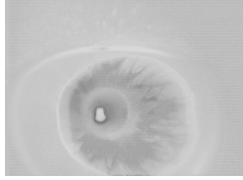

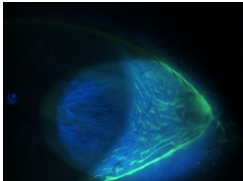
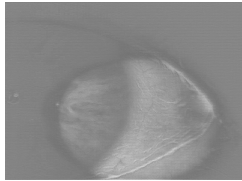
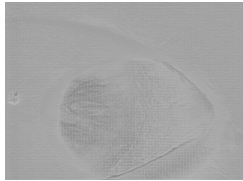

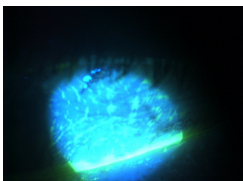

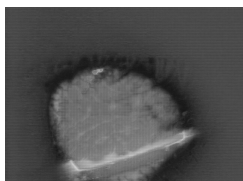

Frame	T ² Net-Vanilla	T ² Net-Full	DenseDepth	IrisDepth
				
				
				
				

Table 4.2 – Depth map predictions part II.

Frame	Depth from videos in the wild	Struct2Depth	Endo-SfMLearner
			
			
			

4.2 Conclusion

In conclusion, our experiments with existing approaches validated our concerns with our data. We know that depth is difficult to estimate from single images, and the motion in the videos is not enough to allow for optical flow training. When looking at the baselines we tested, qualitatively, we could not assess our depth maps. Therefore we move on with two of the methods mentioned in Section 4.1.2 that had obtained state-of-the-art results. In this section these self-supervised methods joined both depth and ego-motion learning, and we want to qualitatively assess their ability to learn and estimate the camera's position for our data. We continue with more in-depth investigations in Chapter 5 with implementations of these baselines in order to obtain qualitative results and further analyze their significance. We also decided to focus more on solving for the disturbances and distinguish the learning for our problematic.

PROPOSED METHOD

“Research is to see what everybody else has seen and to think what nobody else has thought.”

— *Albert Szent-Gyorgyi*

5.1	Introduction	74
5.2	Semantic Segmentation	74
5.3	Baseline assessment	78
	5.3.1 Inverse Warping	80
	5.3.2 Losses	81
	5.3.3 Experiments & Results	84
5.4	Semantic reconstruction loss	87
	5.4.1 Training	89
	5.4.2 Results	90
5.5	Shape Fitting	94
	5.5.1 Sphere fitting loss	95
	5.5.2 Training	98
	5.5.3 Results	98
5.6	Conclusion	100

This chapter explores the development of our proposed method for depth and egomotion prediction from slit-lamp eye examination videos. We focus on the reasons that cause the failure of the baseline methods on our data, mainly hindering the optimization. Starting with simple modifications to answer these problems, we obtain a stable training as a baseline to compare our proposed method to. Next, through two novel developments we obtain a successful, fully self-supervised image registration algorithm.

5.1 Introduction

Following our experiments presented in Chapter 4, moving forward we wanted to add semantic segmentation to extend two of the methods we detailed in chapter 4: Depth from videos in the wild and Struct2depth [183, 188].

These baselines are able to incorporate both depth and camera motion estimation to better predict the environment captured by the camera. The incorporation of depth image-based rendering is the key to linking both estimations. This also allows the use of projective geometry and a self-supervised method of estimating camera motion with the least amount of external input data. This is beneficial as we have depth information simultaneously that we also think we can include in our improvements.

The main drawbacks of all these methods is their strong dependence on the photometric loss. The main guidance for models to learn in these baseline models is that color appears similar from any camera view-point. For us, this is a drawback as we are aware that photometric consistency is lacking in our data.

We measure the accuracy of pose estimation that is learned and predicted through the registration error. Therefore registration error is the first thing we wanted to address and improve in our proposed approach. We investigate three main limitations of the baselines:

1. Heterogeneity of information fed to the model (in section 5.4)
2. Training disturbances due to specular reflections (in section 5.4)
3. Overlooking the complete structure of the eye (in section 5.5)

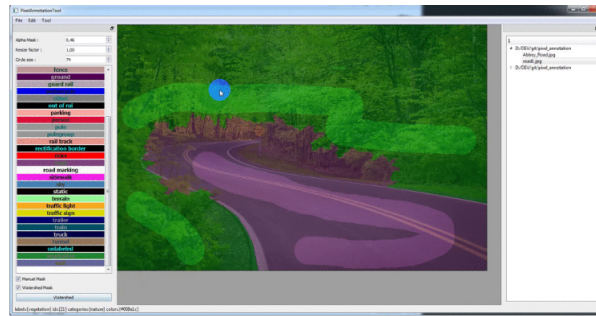
We addressed these three key issues by first alternating the use of the semantic segmentations in the pre-processing. We introduce a new loss that replaces the photometric loss, the semantic reconstruction loss, as a primary supervision signal and thus rids us of the specular reflections interfering with the training. We also benefit from the eye anatomy and look at shape fitting, expanding this to fit our problematic giving us an original loss, sphere fitting loss. In this chapter will take a look at these evolution in details. We start by presenting the preliminary task of semantic segmentation, used in all the proposed improvements.

5.2 Semantic Segmentation

Semantic segmentation is commonly implemented in visual odometry as it allows for better scene understanding. This is particularly relevant in applications where we don't have any sensor information, unlike autonomous driving for instance. To give our method additional information, we implemented state-of-the-art models for semantic segmentation.

We manually segmented 200 randomly selected frames. This was done using the PixelAnnotationTool [190] which is a software that helps manual annotations of images. It uses the algorithm 'watershed marked' of OpenCV and the user has to provide a marker with the brush that produces a segmentation. This can also be corrected and refined if need be.

Figure 5.1 – PixelAnnotationTool [190].



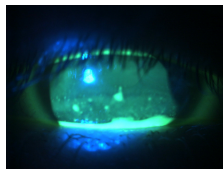

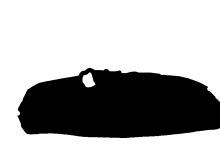

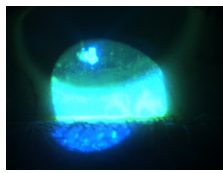
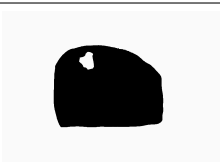


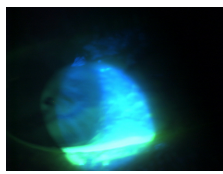


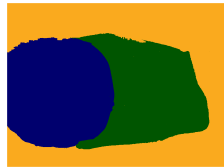
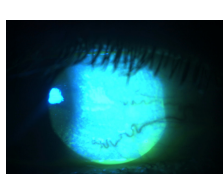

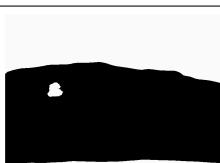

Segmentation for disturbances

We first created manual annotations for binary masks keeping only the region of interest to be: the eye parts illuminated by the light. We ignored any over-illuminated areas, light reflections, eyelashes and eyelids. Table 5.1 shows what we named 'Mask for disturbances'; disturbances that we ignore in white, and in black the areas of the image we keep. We also improved these masks to define a criterion for selecting frames. We decided to exclude the light used during the examination as a reference and focus on the visible part of the eye. With these new masks titled 'Binary mask ROI', all the visible ocular surface is segmented, seen in black, although we continued to ignore the light reflections as can be seen in Table 5.1 in white along with the eyelid. Only frames with a minimum of 40% of non-zero pixels are further analyzed and utilised for training. This rule was implemented as a pre-processing step which allowed us to ensure only viable frames that included a visible ocular surface we kept.

Segmentation for ROI Our final annotations are semantic segmentation with three regions defined: eyelid + eyelashes, cornea and conjunctiva. Following the same manual annotation method we obtain masks with three distinct values for each region: eyelid & eyelashes = 0, cornea = 1 and conjunctiva = 2. For this case we decided to no longer highlight the disturbances but benefit from knowing the anatomy of the eye. We wanted to include this information into our method to help us imitate how the Oxford grading is done by the ophthalmologist. As we mentioned previously, in Chapter 3 the Oxford grading scale consists of three 0-5 grades for each of the sections: cornea, nasal conjunctiva and temporal conjunctiva. By distinguishing these areas in our frames we could treat the cornea and conjunctiva separately for future tasks, as they are our regions of interest (ROI). These masks shown in Table 5.1. 'Mask for ROI' were also used in a pre-processing step. By setting a condition that at least 40% of the ROI (the eye)

should be visible, we retain only frames with valuable information.

Table 5.1 – Semantic segmentation training examples.

No.	Frame	Mask for disturbances	Binary mask ROI	Mask for ROI
1				
2				
3				
4				

* Mask for disturbances ; ignored area = white, ROI = black.

* Binary mask ROI ; ignored area = white, ROI = black.

* Mask for ROI ; eyelashes & eyelid = yellow, green = conjunctiva, blue = cornea.

Training set-up

Experiments

In order to obtain our fully trained model that could predict any of the three semantic segmentations we annotated in Table 5.1, we utilised the library Segmentation Models [191]. Segmentation models is a PyTorch Module that can be easily incorporated to any code. It facilitates the creation of models and contains pre-trained weights for faster and better convergence.

We tested two of the nine available architectures: U-Net [192] and FPN (Feature Pyramid Network) [193]. We evaluated our models using the Sørensen–Dice coefficient which is also the metric of training. More commonly referred to as the Dice coefficient (F1 Score), it is used to compare the pixel-wise agreement between a predicted segmentation and its corresponding ground truth. It ranges from 0 to 1, with 1 signifying the greatest similarity between predicted and ground truth.

$$DiceScore = \frac{2 * |XY|}{|X| + |Y|} \quad (5.1)$$

where X is the predicted set of pixels and Y is the ground truth.

Given our small manually annotated set for the training, we included data augmentation. This technique is used to increase data amount by applying transformations and creating slightly modified versions of the data. We performed online augmentation, using random transformations from the Augmentor python package [194]. Online augmentation is a preferred method as it doesn't require pre-processing and saves memory. We applied the following methods of distortion: rotation, flipping left to right, flipping top to bottom, random zoom.

Figure 5.2 shows some of the examples of the data augmentation step. Details of the training and results are listed in Table 5.2.

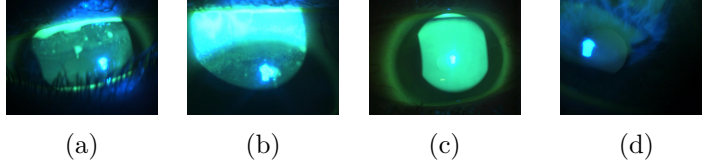
For the following task the experiments used a combination of different models and encoders with a fixed set of hyper-parameters. The hyper-parameters used are:

1. Model pre-trained: ImageNet.
2. Loss function: Cross entropy.
3. Optimizer: Adam.
4. Learning rate: 10^{-6} .
5. Batch size: 32.
6. Epochs: 300.

Table 5.2 – Dice score of all experiments.

Mask for disturbances			Binary mask ROI			Mask for ROI		
Model	Encoder	Dice Score	Model	Encoder	Dice Score	Model	Encoder	Dice Score
U-Net	resnet50	0.88	U-Net	resnet50	0.83	U-Net	resnet50	0.96
FPN	resnet50	0.83	FPN	resnet50	0.85	FPN	resnet50	0.87
U-Net	EfficientNet-B3	0.87	U-Net	EfficientNet-B3	0.82	U-Net	EfficientNet-B3	0.92
FPN	EfficientNet-B3	0.90	FPN	EfficientNet-B3	0.85	FPN	EfficientNet-B3	0.95

Figure 5.2 – Frame augmentation examples.



5.3 Baseline assessment

As an extension of the two methods we implemented a hybrid that took into account the main architectures of the proposed methods and the main losses [183, 188]. We first detail the main elements from [183, 188] we want to maintain that make up our hybrid implementation. This includes the training framework and the weights attributed to three of the main losses. We demonstrate below how the semantic segmentations we obtained are utilised in the baseline framework and present different experiments and their limitations. We investigate the reasons for these limitations and to show that the assumptions made in these baseline approaches are defied in our case, and do not align with our problematic.

We start with a the baseline method (hybrid of [183, 188]) and in order to train it we first set-up the input framework (which is similar for both methods). Although they detect object-motion by using masks, our first experiment assumes no occlusion in our examinations first. Given that our main reservations with these methods remains the quality of our images, we wanted to ensure and test the limitations of what valuable information can be learned. We implemented the framework with a slightly different approach excluding object-motion detection, but rather guide the egomotion training to focus on certain regions and ignore those that were poorly lit.

The method is made up of two convolutional neural networks (CNN); one predicts depth from a single image, while the other uses two images to predict egomotion, object motion field in relation to the scene, and camera intrinsics. The first CNN is a UNet architecture with a ResNet 18 base. It has a softplus activation ($z = \log(1 + e^\ell)$) to convert the logits (ℓ) to depth (z). The second CNN for the egomotion estimation is inspired by FlowNet [195]. The final output, 3 channels, each predict the global rotation angles (r_0) and translation vector (t_0). The complete method gives a depth map prediction and the estimation of the scene movement with respect to the camera. In this experiment we also used the masks as done in [183] to ignore certain regions when training the egomotion CNN. When predicting egomotion we use the masks for disturbances $m(x, y)$ as shown in Table 5.1 to ignore any disturbances. Following [183]’s implementation, this was added as follows for a warping of pair of frames $\hat{I}_s - I_t$:

$$I_s \sim K\hat{T}_{t \rightarrow s}\hat{D}_t(I_t)K^{-1}I_t \quad (5.2)$$

Table 5.3 – Specification of DispNet architecture [164].

Name	Kernel	Str.	Ch I/O	InpRes	OutRes	Input
conv1	7×7	2	6/64	768×384	384×192	Images
conv2	5×5	2	64/128	384×192	192×96	conv1
conv3a	5×5	2	128/256	192×96	96×48	conv2
conv3b	3×3	1	256/256	96×48	96×48	conv3a
conv4a	3×3	2	256/512	96×48	48×24	conv3b
conv4b	3×3	1	512/512	48×24	48×24	conv4a
conv5a	3×3	2	512/512	48×24	24×12	conv4b
conv5b	3×3	1	512/512	24×12	24×12	conv5a
conv6a	3×3	2	512/1024	24×12	12×6	conv5b
conv6b	3×3	1	1024/1024	12×6	12×6	conv6a
pr6+loss6	3×3	1	1024/1	12×6	12×6	conv6b
upconv5	4×4	2	1024/512	12×6	24×12	conv6b
iconv5	3×3	1	1025/512	24×12	24×12	upconv5+pr6+conv5b
pr5+loss5	3×3	1	512/1	24×12	24×12	iconv5
upconv4	4×4	2	512/256	24×12	48×24	iconv5
iconv4	3×3	1	769/256	48×24	48×24	upconv4+pr5+conv4b
pr4+loss4	3×3	1	256/1	48×24	48×24	iconv4
upconv3	4×4	2	256/128	48×24	96×48	iconv4
iconv3	3×3	1	385/128	96×48	96×48	upconv3+pr4+conv3b
pr3+loss3	3×3	1	128/1	96×48	96×48	iconv3
upconv2	4×4	2	128/64	96×48	192×96	iconv3
iconv2	3×3	1	193/64	192×96	192×96	upconv2+pr3+conv2
pr2+loss2	3×3	1	64/1	192×96	192×96	iconv2
upconv1	4×4	2	64/32	192×96	384×192	iconv2
iconv1	3×3	1	97/32	384×192	384×192	upconv1+pr2+conv1
pr1+loss1	3×3	1	32/1	384×192	384×192	iconv1

where I_s are pixels in the source frame, I_t are pixels in the target frame

$$t(x, y) = t_0 + m(x, y)\delta t(x, y). \quad (5.3)$$

where t_0 is the translation vector and $m(x, y)$ equals one at pixels that could belong to disturbances and zero otherwise.

Architecture of DispNet [164] is detailed in Table 5.3, and of FlowNet [195] in 5.4.

Table 5.4 – Specification of FlowNet architecture [195].

Name	Kernel	Str.	Ch I/O	InpRes	OutRes	Input
conv1	7×7	2	6/64	384×512	192×256	Images
conv2	5×5	2	64/128	192×256	96×128	conv1
conv3	5×5	2	128/256	96×128	48×64	conv2
conv3 ₁	3×3	2	256/256	48×64	48×64	conv3
conv4	3×3	2	256/512	48×64	24×32	conv3 ₁
conv4 ₁	3×3	2	512/512	24×32	24×32	conv4
conv5	3×3	2	512/512	24×32	12×16	conv4 ₁
conv5 ₁	3×3	2	512/512	24×32	12×16	conv5
conv6	3×3	2	512/1024	12×16	6×8	conv5 ₁

5.3.1 Inverse Warping

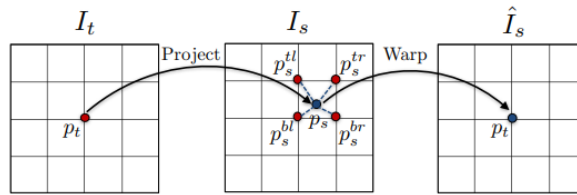
The main supervision signal that ties in Equation 5.2 and the key is that the model learns through the difference between the reconstructed target image using source pixels and comparing them to the original target frame. This is named inverse warping or backward warp, which has been proven to be much easier to optimize than forward warping [196]. In forward warping every point in the source image is transformed to obtain a new warped target image. This method results in holes and splattering, which requires normalization. The differentiable process was introduced by [176], and illustrated in the Figure 5.3. Based on predicted depth and camera motion we reconstruct I_t by sampling pixels from I_s . All points p_t are projected onto source view and then using bilinear interpolation we obtain the value of final warped image \hat{I}_s . The sampling mechanism is differentiable bilinear sampling proposed in the spatial transformer networks [197]. The method linearly interpolates using the values of the 4-pixel neighbors to approximate the values of $\hat{I}_s(p_t)$, shown in Equation 5.4. By using projective geometry to warp pixels we are able to include depth and camera pose estimation and use this differentiable depth image-based rendering as a supervision signal.

$$\hat{I}_s(p_t) = I_s(p_s) = \sum_{i \in \{t,b\}, j \in \{l,r\}} w^{ij} I_s(p_s^{ij}) \quad (5.4)$$

where p_t are pixels that belong to the target frame I_t , p_s to the source frame I_s , and \hat{I}_s the warped frame,

where w^{ij} is linearly proportional to the spatial proximity between p_s and p_s^{ij} , and $\sum_{i,j} w^{ij} = 1$,

Figure 5.3 – Differentiable depth image-based warping [176]



and t=top, b=bottom, l=left, r=right.

These mechanisms also include implicit assumptions and therefore limitations to the model learning:

1. Static scene with no moving objects.
2. No occlusion between camera views.
3. A Lambertian surface for a meaningful photo-consistency error. Where the surface appears uniformly bright from all angles and has the lambertian reflectance property.

A violation of these assumptions inhibits the training and also could corrupt the overall gradient. We assume that our videos are comprised of a static eye with the camera moving (left, right) during the examination. Therefore our data complies with the first assumption, as for the second we use our semantic segmentation to exclude the only occlusion we have: eyelid. Our predicted masks allow us to pre-process the dataset and only keep frames where the eye is visible and so removing any blinking frames, or half open eye frames. Lastly, the eye is not a Lambertian surface and a main disturbance is the specular reflections. The cornea is also a transparent part of the eye making this another major disturbance when attempting to calculate an error based on consistency in the surface pixel value.

5.3.2 Losses

There are various losses, some for either CNNs specifically and a common main supervision signal that we previously mentioned. This equation shows the relation between two adjacent video frames using a depth map and the camera intrinsic parameters:

$$z'p' = KRK^{-1}zp + Kt \quad (5.5)$$

where p and p' are pixel coordinates in homogeneous form before and after the transformation represented by the rotation matrix R and the translation vector t . z and z' are the respective depths.

Notations

- Triplet frames: $I : [I_{t-n}, I_t, I_{t+n}]$
- Step between frames: n
- Pair of frame from a triplet : I_i, I_j
- Depth map: D
- Egomotion transform of $i \rightarrow j$: $E_{i \rightarrow j}$

The losses used to obtain the warped image, using a differentiable image warping operator $\phi(I_i, D_j, E_{i \rightarrow j}) \rightarrow \hat{I}_{i \rightarrow j}$, make up the total loss for every pair of frames $\hat{I}_{i \rightarrow j} - I_j$. Two losses are used as image reconstruction guidance; photometric loss & SSIM, while depth smoothness is focused on smoothing the predicted depth maps. Both depth and pose estimations are optimized through the image reconstruction losses. The first, photometric loss, is the difference between corresponding pixels. This simple L1 loss is presented below Equation 5.6.

Photometric loss ($_{\text{RECON}}$) frames are used as input to compare this reconstructed image $\hat{I}_{i \rightarrow j}$ to the next frame I_j [188].

$$\mathcal{L}_{recon} = \|\hat{I}_{i \rightarrow j} - I_j\| \quad (5.6)$$

Structural similarity loss ($_{\text{SSIM}}$) is used to assess the quality of the warping [198]. The introduction of structured similarity (SSIM) [198] included an evaluation of the quality of the predicted image as well. SSIM index measures the similarity between images in terms of luminance, contrast and structural information. We measure this between $\hat{I}_{i \rightarrow j}$ and I_j . Luminance of an image signal is estimated by mean intensity 5.7 & luminance of two images is then compared by 5.8:

$$\mu_x =; \frac{1}{N} \sum_{i=1}^N x_i \quad (5.7)$$

$$l(x, y) = \frac{2\mu_x\mu_y + l_1}{\mu_x^2 + \mu_y^2 + l_1} \quad (5.8)$$

Contrast is measured by difference of the luminance between objects in the field of view. This is done by calculating the standard deviation of the image signal 5.9, and the contrast similarity is calculated by 5.10.

$$\sigma_x = \left(\frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)^2 \right)^{\frac{1}{2}} \quad (5.9)$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + l_2}{\sigma_x^2 + \sigma_y^2 + l_2} \quad (5.10)$$

Strong inter-dependencies between relatively near pixels are used to represent structural information. Image signals are projected as unit vectors on hyperplanes defined by 5.11, the signals are normalized by subtracting the mean intensities and then dividing by respective standard deviation. The structural data is related to these unit vectors 5.12, which then gives the correlation between the two windows 5.13.

$$\sum_{i=1}^N x = 0 \quad (5.11)$$

$$s(x, y) = \frac{\sigma_{xy} + l_3}{\sigma_x\sigma_y + l_3} \quad (5.12)$$

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y) \quad (5.13)$$

$$SSIM(\hat{I}_{i \rightarrow j}, I_j) = l(\hat{I}_{i \rightarrow j}, I_j) * c(\hat{I}_{i \rightarrow j}, I_j) * s(\hat{I}_{i \rightarrow j}, I_j) \quad (5.14)$$

$$SSIM(\hat{I}_{i \rightarrow j}, I_j) = \frac{(2\mu_{\hat{I}_{i \rightarrow j}}\mu_{I_j} + l_1)(2\sigma_{\hat{I}_{i \rightarrow j}I_j} + l_2)}{(\mu_{\hat{I}_{i \rightarrow j}}^2 + \mu_{I_j}^2 + l_1)(\sigma_{\hat{I}_{i \rightarrow j}}^2 + \sigma_{I_j}^2 + l_2)} \quad (5.15)$$

where $\mu_{\hat{I}_{i \rightarrow j}}, \mu_{I_j}$ are the average, $\sigma_{\hat{I}_{i \rightarrow j}}^2, \sigma_{I_j}^2$ the variance, and $\sigma_{\hat{I}_{i \rightarrow j}I_j}$ covariance of $\hat{I}_{i \rightarrow j}, I_j$, where $l_1 = (k_1L)^2$, $l_2 = (k_2L)^2$, $l_3 = l_2/2$, L is the dynamic range of the pixel-values, $k_1 = 0.01$, $k_2 = 0.03$.

The SSIM loss, which is used as part of the objective function, is given by:

$$\mathcal{L}_{SSIM} = 1 - SSIM(\hat{I}_{i \rightarrow j}, I_j) \quad (5.16)$$

Together these make up the main loss for the image reconstruction task as has been used in various methods [174, 199].

To address the gradient-locality in motion estimation, a smoothness term is introduced to avoid discontinuity of the learned depth maps in regions with low-texture. The edge-aware depth smoothness loss used in [174] uses image gradient to weigh the depth gradient.

Depth smoothness (DS) encourages smoothness by penalizing depth discontinuity if the image shows continuity in the same area [173].

$$\mathcal{L}_{DS} = |\nabla_x D_i| e^{-\nabla_x I_i} + |\nabla_y D_i| e^{-\nabla_y I_i} \quad (5.17)$$

where ∇_x, ∇_y are image gradients in the horizontal and vertical direction, respectively, ∇ denotes the 2D differential operator, and $|\cdot|$ is the element-wise absolute value.

The total loss is made up of a sum of the losses mentioned where each is multiplied by a weight $\alpha_a, \alpha_b, \alpha_c$.

$$\mathcal{L}_{total} = \alpha_a \mathcal{L}_{recon} + \alpha_b \mathcal{L}_{SSIM} + \alpha_c \mathcal{L}_{DS} \quad (5.18)$$

Following implementations [183, 188] the weights were the following: $\alpha_a = 0.85$, $\alpha_b = 0.15$, and $\alpha_c = 0.04$.

5.3.3 Experiments & Results

We trained the models with the inputs shown in Table 5.7. Both experiments used a set of triplet frames as input : $I : [I_{t-n}, I_t, I_{t+n}]$.

1. Depth: trains with single frame from the triplet and produces a depth map.

$$D_i = \theta(I_i), \theta : \mathbb{R}^{(H \times W \times 3)} \rightarrow \mathbb{R}^{(H \times W)} \quad (5.19)$$

2. Egomotion: the network takes three frames (ex. $[I_{t-n}, I_t, I_{t+n}]$) and predicts transformations simultaneously

$$\psi_E(I_{i-n}, I_i) = (t_{x_1}, t_{y_1}, t_{z_1}, r_{x_1}, r_{y_1}, r_{z_1}) \quad (5.20)$$

$$\psi_E(I_i, I_{i+n}) = (t_{x_2}, t_{y_2}, t_{z_2}, r_{x_2}, r_{y_2}, r_{z_2}) \quad (5.21)$$

The data was split into ; 35 eyes for the train, 39 for the validation and 14 for the test. All eyes from the same patient were assigned to the same set.

We tested the hybrid implementation by first training with masks (experiment inputs 'a') that take our full frames into account and therefore ignore nothing. Followed by an experiment with inputs 'b' where we utilised our masks for disturbances to help the egomotion CNN focus only on the well lit areas. These are shown in Table 5.5. For both experiments we kept the same set-up. The pre-trained model we used to initiate the training was trained on the Cityscape dataset. With this set-up we noticed that the loss diverged constantly. Training with different parameters, and randomly initialised weights unfortunately did not help either. Neither set-ups or changes helped the models learn any valuable information. These changes include the mask changes as well as changing the weights of the losses or using randomly initialized weights. The main issue was also the exploding gradient after very few epochs where the loss diverged to $\mathcal{L}_{total} \rightarrow \infty$. The optimization problem proved to be too difficult. Table 5.6 demonstrates two

loss plots of the experiments that constantly required a form of 'restart training'. With no sign of convergence the training optimization seemed to be unsuccessful given the set-up we were using.

We concluded that if we took into the full image (inputs 'a'), or masks for disturbances (inputs 'b') as shown in Table 5.5 for the equation 5.3, there was no improvement in the training. The difficulty we faced in these experiments demonstrated that the main losses do not work for our data and problematic. The optimization difficulty surpassed the constrains implemented when using our dataset. Not only were the losses not enough, we probably did not make the best use of the masks. Attempting to change the object-motion detection to ignore disruptions seemed to have limited effect in the current implementation.

As we had theorised, our data defies the assumptions made and the main components that seemed to be common in several baseline methods were inadequate.

Table 5.5 – Training input examples.

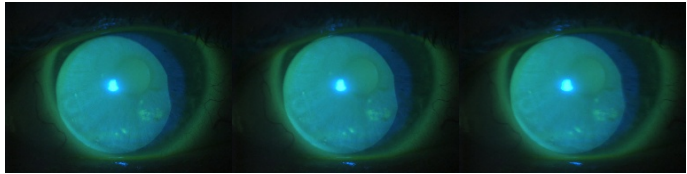
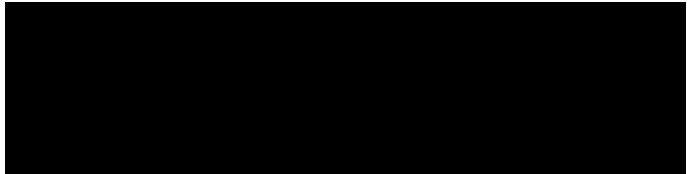

Exp. Index	Inputs	Example
a & b	Frames	
a	Mask (no regions ignored)	
b	Mask for disturbances	
a & b	Intrinsic parameters	$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1.0 \end{bmatrix} = \begin{bmatrix} 3758.9 & 0 & 138.8 \\ 0 & 3758.9 & 85.4 \\ 0 & 0 & 1.0 \end{bmatrix} \quad (5.22)$

Table 5.6 – Experiment loss & training details.

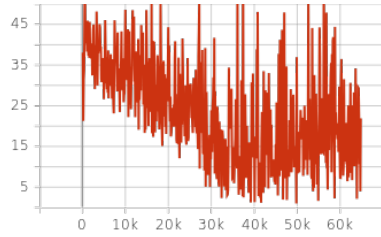
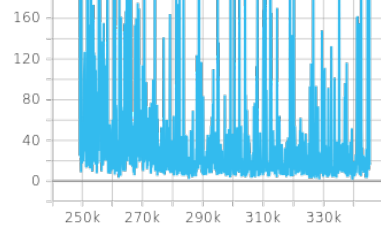
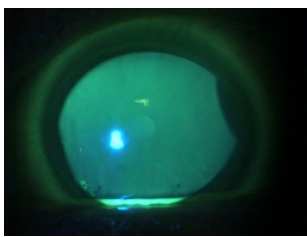
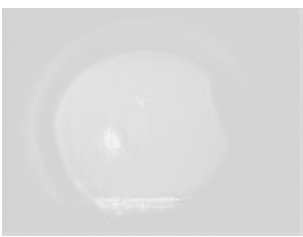
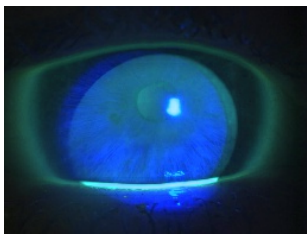
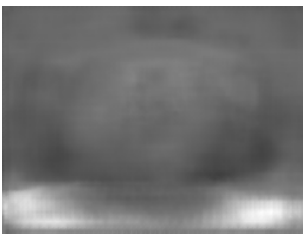
Exp. Index	Loss	Training Time	Loss plot
a	21.91	21hrs	
b	19.93	72hrs	

Table 5.7 – Experiment results.

Exp. Index	Frame	Depth Prediction
a		
b		

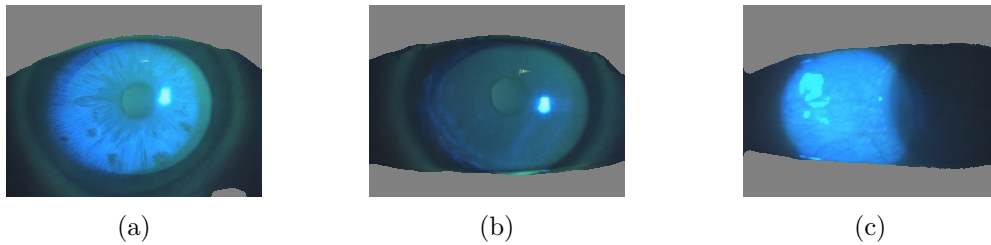
5.4 Semantic reconstruction loss

Our previous experiments demonstrated the difficulty of training when relying on the photometric loss mainly, as it was given the highest weight to guide the depth egomotion learning. We decided to retain the use of the masks although we didn't see any improvement in the way

we used them in our previous experiments. Instead we wanted to incorporate them to help us minimize our model taking in any trivial information. This is mainly the eyelid areas present in our frames, which doesn't contain any information of good quality. The second point that we were able to prove in the previous experiments is that the photometric loss can not be the highest weighted source of guidance to train the model, and therefore the strongest penalisation for the models. Our hypothesis that inconsistency present in our frames, which is due to the lighting, was confirmed by the experiments in Section 5.3.3. The constant movement of the light, being that it is attached to the camera, wasn't a minor but a major disturbance during the training. This also includes the artifacts of reflections and overly exposed areas due to the light.

Before detailing the semantic reconstruction loss, a set of steps were taken that led to its proposal that we wish to discuss. We started by addressing the removal of non-essential region ; eyelid using the binary masks we named 'Binary mask ROI' shown in Table 5.1. We used these binary masks ROI and multiplied them by the frames as a pre-processing step for training. We then obtained frames that contain only the eye, examples shown in 5.4. This was easily incorporated into our code as a variable which would allow us to easily interchange the inputs for training. Another pre-processing step that we incorporated is the choice of a step of n frames as we believe the motion between two consecutive frames was limited. The baseline approaches do not utilise this pre-processing and rely on consecutive frames when creating the dataset.

Figure 5.4 – Processed frames with binary mask ROI.



The interpretations of the previous results led us to focus on the main loss, the photometric $\text{loss}_{(\text{RECON})}$, which had the highest weight in the total loss in our baseline assessment 5.3. The photometric loss compared the reconstructed image $\hat{I}_{i \rightarrow j}$ to the next frame I_j . This is the main supervision signal that utilises both predictions from the Depth and Egomotion CNNs (see Eq.5.6).

To enable the Egomotion CNN to learn valuable camera motion information we had to find a way to get rid of the light seen on the frames. We wanted to be able to compare the estimated motion between frames without relying solely on the photometric consistency. Using the semantic segmentations we introduced earlier as 'Mask for ROI' in Table 5.1, referred to as m_{ROI} , we had three binary masks that identified three regions in our frames ; eyelid, cornea and

sclera, detailed in Eq 5.23. With each of these regions' pixels in the mask ROI having a value of ; 0, 1 and 2 respectfully we implemented a new semantic reconstruction loss \mathcal{L}_{SRL} .

$$\begin{cases} m_{eyelid}(x, y) = 1 & \iff m_{ROI}(x, y) = 0 \\ m_{cornea}(x, y) = 1 & \iff m_{ROI}(x, y) = 1 \\ m_{sclera}(x, y) = 1 & \iff m_{ROI}(x, y) = 2 \end{cases} \quad (5.23)$$

The loss is based on Eq.5.6, in which we still utilise the predicted depth and camera motion. The inverse warping step remains the same and the predicted transformation matrix is now used to warp the mask ROI instead. Although the model takes in our processed frames Fig5.4 to learn, we measure the error using the masks instead of the raw frame input, as done in the baselines [183, 188] using 5.6. This removes entirely the need for color consistency which is how both the depth and egomotion models were previously penalised via the \mathcal{L}_{recon} loss (Eq.5.6). We also calculate this loss per region which allows us to precisely measure the error of reconstruction for both the cornea and sclera.

Semantic reconstruction loss (\mathcal{L}_{SRL}) is the main supervision signal.

$$\mathcal{L}_{SRL} = \|\hat{m}_{cornea_{i \rightarrow j}} - m_{cornea_j}\| + \|\hat{m}_{sclera_{i \rightarrow j}} - m_{sclera_j}\| \quad (5.24)$$

where $\hat{m}_{i \rightarrow j}$ is the reconstructed mask ROI, m_j the target mask ROI.

This new addition to training was set as our main loss, while maintaining \mathcal{L}_{recon} 5.6, \mathcal{L}_{SSIM} 5.16, \mathcal{L}_{DS} 5.17, that continue to be calculated with frame inputs.

The total loss is now:

$$\mathcal{L}_{total} = \alpha_a \mathcal{L}_{recon} + \alpha_b \mathcal{L}_{SSIM} + \alpha_c \mathcal{L}_{DS} + \alpha_d \mathcal{L}_{SRL} \quad (5.25)$$

where $\alpha_a = 0.85$, $\alpha_b = 0.15$, $\alpha_c = 0.04$., and $\alpha_d = 1$.

5.4.1 Training

The overall training setup included fixed values for the weights, the learning rate = 0.0002, and the batch size = 8. We trained all models for 200 epochs and used the 'best model' for inference. The best model is defined as the model with the lowest total loss on the training set. Variables that were investigated to allow us to assess our newly modified approach is the inclusion of the frame step, the original photometric loss \mathcal{L}_{recon} . Lastly, we detail which dataset was used for training and which was used for pre-training. Tables 5.8,5.9,5.10 detail the experiments and results. The notation used are the following:

- \mathcal{L}_{recon} = Photometric loss

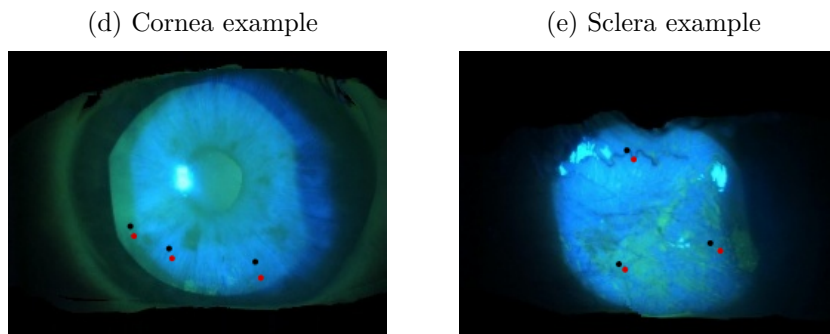
- \mathcal{L}_{SRL} = Semantic reconstruction loss
- Dataset 'P' = PEPPS dataset
- Dataset 'O' = Original dataset
- Pre-train Dataset 'C' = model trained on the Cityscapes dataset [172]

5.4.2 Results

For the qualitative results we used the annotations we detailed in Chapter 3. One way we could ensure the egomotion estimation is correct is mark the coordinates of distinct points in various frames. These points were annotated on static parts of the eye. The annotated punctate dots (damaged area) were on the surface of the eye, and visible veins on the sclera. To visualise the accuracy of our predictions we warp a source frame into a target frame and track the marked points.

Figure 5.5 shows an example of this qualitative evaluation method that includes the main supervision signal that trains both the depth and egomotion CNN. Our evaluation uses the inverse warping which first requires the depth map \hat{D}_t prediction of the target frame, and then the egomotion of $source \rightarrow target$.

Figure 5.5 – Registration evaluation.



We present our first results in Table 5.8. As mentioned we had tested changing various variables. Focusing on the Mean Euclidian distance (px,%) to evaluate these experiments we notice an improvement of $\approx 21px, 1.29\%$ when comparing the baseline to our proposed method which includes the semantic reconstruction loss, \mathcal{L}_{SRL} (see Eq. 5.24). To review these results in more depth, we first started with the initial approach, detailed in 5.3, which uses the total loss (see Eq. 5.18), and consecutive frames for input, or a frame step of $n = 1$. The results for this setup up is referred to as Exp. No. A1 in Table 5.8. This gave us our baseline results where we trained with the database 'P', and used the model pre-trained on database 'C'.

In order to verify that there was a lack of motion we added a frame step of $n = 10$ and obtained an improvement of $\approx 1.7px, 0.1\%$, referred to as Exp. No. A2 in Table 5.8. We then decided to continue our training with a frame step of $n = 10$, as it showed some promise. Adding our newly proposed semantic reconstruction loss \mathcal{L}_{SRL} , continued to improve results. The results of this setup, which included both a frame step of $n = 10$, and the semantic reconstruction loss \mathcal{L}_{SRL} is referred to as Exp. No. A3 in Table 5.8. Given that we had the database 'O', we also wanted to test the setup used for Exp. No. A3 and train the models first with database 'O' to be able to obtain a pre-trained model before moving forward with dataset 'P'. The results validated that although the model can be exposed to a dataset of lesser quality, dataset 'O', it resulted in a better training set-up than using a model pre-trained on a completely different set of images from database 'C'.

Table 5.8 – Experiment details and results A.

Exp. No.	Frame step	DepthNet Input	EgoNet Input	\mathcal{L}_{SRL}	\mathcal{L}_{Recon}	Mean Euclidian (px)	Mean Euclidan (%)	Training dataset	Pre-train dataset
A1	n=1	Frames	Frames	No	Yes	33.67	2.10	P	C
A2	n=10	Frames	Frames	No	Yes	31.96	2.00	P	C
A3	n=10	Frames	Frames	Yes	Yes	27.07	1.69	P	C
A4	n=10	Frames	Frames	Yes	Yes	22.48	1.4	P	O
Grader errors	-	-	-	-	-	5.30	0.33	-	-

Table 5.8 summarises the first set of experiments, and includes the grader error. The grader error was measured as a mean of all the annotations performed by experts, as described in 3.5. We deduced from these results that we wanted to maintain a frame step $\neq 1$ and the model that was pre-trained using dataset 'O'. As we mentioned previously we wanted to minimize any irrelevant information, which includes the eyelid and eyelashes, being fed to the models when training so we changed our input to the binary frames, as shown in Figure 5.4. Our new main loss \mathcal{L}_{SRL} was promising, so we wanted to test training the egomotion CNN with only the 'Mask for ROI' as input, which is what is used to calculate the \mathcal{L}_{SRL} and shown in Fig. 5.1. These are the inputs we set for Experiments B, along with keeping \mathcal{L}_{DS} and \mathcal{L}_{SSIM} .

For these set of experiments B, we continued to see improvements when keeping the photometric loss \mathcal{L}_{recon} of $\approx 5.7px, 0.36\%$, versus using only our proposed \mathcal{L}_{SRL} along with \mathcal{L}_{DS} and \mathcal{L}_{SSIM} . We increased the frame step from $n = 10$ to $n = 30$ but with these inputs, which seemed to have a negative effect. Experiments B3,B4 with $n = 30$ proved that there was a limitation to the motion between frames that we need to respect. Although B3,B4 remained an improvement to experiments without the \mathcal{L}_{SRL} (A1,A2), we deduced that $n = 10$ was an ideal frame step to train with.

Table 5.9 – Experiment details and results B.

Exp. No.	Frame step	DepthNet Input	EgoNet Input	\mathcal{L}_{SRL}	\mathcal{L}_{Recon}	Mean Euclidian (px)	Mean Euclidian (%)	Training dataset	Pre-train dataset
B1	n=10	Frames Binary	Segmentation	Yes	No	22.61	1.41	P	O
B2	n=10	Frames Binary	Segmentation	Yes	Yes	17.64	1.10	P	O
B3	n=30	Frames Binary	Segmentation	Yes	No	24.67	1.54	P	O
B4	n=30	Frames Binary	Segmentation	Yes	Yes	25.12	1.57	P	O
Grader errors	-	-	-	-	-	5.30	0.33	-	-

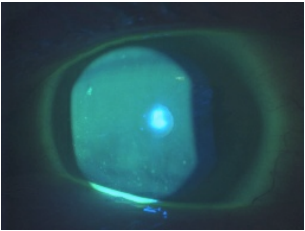
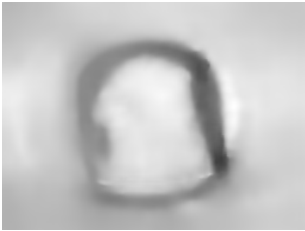
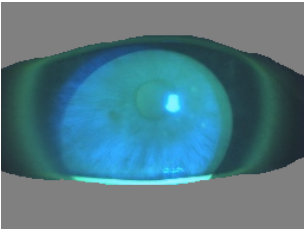
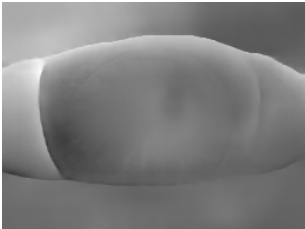
Lastly, the results from experiment B showed that both losses \mathcal{L}_{SRL} and \mathcal{L}_{recon} worked better together. We reverted to maintaining the same input for both CNNs and using the mask ROI only for the \mathcal{L}_{SRL} loss calculation. These results for experiment C are shown below in Table 5.10. Again, keeping the \mathcal{L}_{recon} continued to improve results by $\approx 1.18px, 0.07\%$. The final set of experiments gave us our best results, but with still a large margin when compared to the human error we obtained when annotating our test set.

Table 5.10 – Experiment details and results C.

Exp. No.	Frame step	DepthNet Input	EgoNet Input	\mathcal{L}_{SRL}	\mathcal{L}_{Recon}	Mean Euclidian (px)	Mean Euclidian (%)	Training dataset	Pre-train dataset
C1	n=10	Frames Binary	Frames Binary	Yes	No	14.10	0.88	P	O
C2	n=10	Frames Binary	Frames Binary	Yes	Yes	12.92	0.81	P	O
C3	n=30	Frames Binary	Frames Binary	Yes	No	16.73	1.05	P	O
C4	n=30	Frames Binary	Frames Binary	Yes	Yes	12.98	0.81	P	O
Grader errors	-	-	-	-	-	5.30	0.33	-	-

With the three set of experiments in Tables 5.8, 5.9, 5.10 we were able to finally stabilize training and improve results. We showed with various combinations that the new semantic reconstruction loss always enhanced results. We also saw visual improvements in the depth map estimations, Fig. ??, and this is validated with the fact that depth is utilised for inverse warping and hence the improvement in our results. As we discussed each experiment we concluded that a certain set of parameters worked best for our training: frame step $n = 10$, frame binary for both CNN inputs and using the model pre-trained on the database 'O'. Our best performing method C2 had a difference of $\approx 8.1px, 0.5\%$ Euclidean distance when compared to the grader error. The photometric reconstruction loss \mathcal{L}_{recon} seemed to be less reliable, giving a noisy training with our data. The semantic reconstruction loss displayed a more global and more dependable supervision signal but still lacking in precision. We deduce this given the difference between our best results, Exp. No. C2, and the grader error. We found that the best results is giving a stronger weight to \mathcal{L}_{SRL} , weight = 1, compared to \mathcal{L}_{recon} , weight = 0.85, but maintaining both. We obtain a more stable enforcement of frame consistency with both losses working together.

Table 5.11 – Depth predictions for experiments.

Exp. No.	Frame	Depth Prediction
A1		
C2		

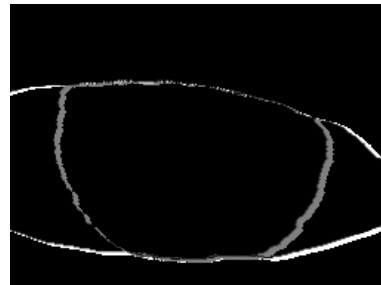
[!ht]

Figure 5.6 – Visualisation of the \mathcal{L}_{SRL} .

(a) Difference between source & target



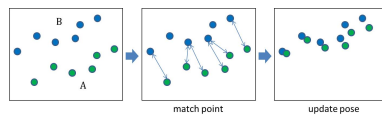
(b) Difference between registered mask ROI & target



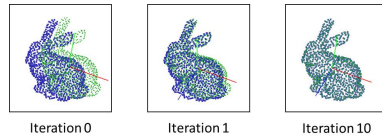
5.5 Shape Fitting

After the introduction of the semantic reconstruction loss we wanted to take advantage of the greatest difference our problematic has compared to others. We have concrete knowledge of what the camera is looking at, the eye. This knowledge can be incorporated into the model to help guide the learned depth to fit a certain shape. The closest to this proposed idea would be the Iterative Closest Point (ICP) algorithm [200]. The general idea in ICP is to constantly revise transformation predictions and minimize an error by measuring the distance between the source point cloud and the target point cloud. This inspired us to add a constraint to our point cloud. Although ICP has also been incorporated as a 3D Point Cloud Alignment Loss in [199], it is not differentiable and so the proposed method only approximated the gradient to allow for back-propagation. ICP also remains a computationally heavy calculation step.

Figure 5.7 – Iterative Closest Point (ICP) algorithm [200].



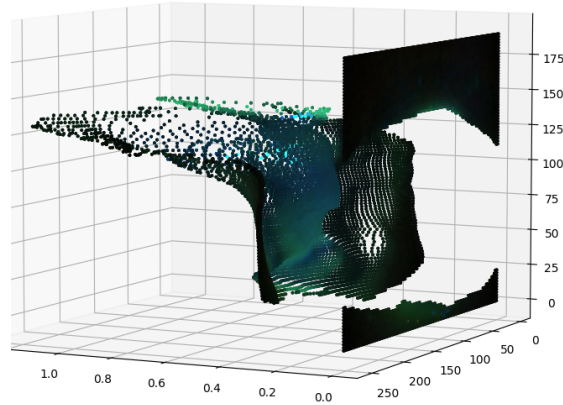
(a) Point matching



(b) ICP Iteration

A visual of a point cloud plot for the best performing proposed method, Exp. No. C2 in section 5.4, is shown below in Fig.5.8. Figure 5.8 is a plot that ignores the eyelid and eyelashes, that are seen as a black plane plotted at depth = 0. By using our semantic segmentations we continue to ignore these areas that hold no valuable information for training. This helped us visualise the errors in the estimations of the current depth and confirmed the need for a stronger constraint to be applied to the depth CNN. The previous section demonstrated that better constraints that fitted our problematic, which was a new loss \mathcal{L}_{SRL} along with the correct set-up improved training and results. Adding the frame step alone allowed us to stabilise training to obtain baseline results, allowing our egomotion CNN to finally learn valuable information. This led us to focus on improving the learning for the depth CNN, as both depth and egomotion estimations are utilised in the key supervision signals \mathcal{L}_{SRL} and \mathcal{L}_{recon} .

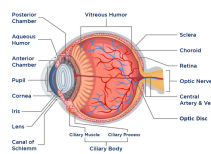
Figure 5.8 – Point cloud plot.



5.5.1 Sphere fitting loss

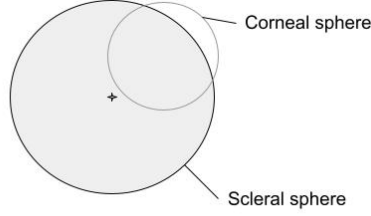
Our final novel loss helps both the depth map and camera pose estimations. This loss is based on modelling the human eye as two intersecting spheres. A sphere for the cornea which is in a larger sphere representing the sclera Fig. 5.10. Studies also show that human adult eye diameter is 24.2 mm (transverse) \times 23.7 mm (sagittal) \times 22.0–24.8 mm [201]. A smaller anterior transparent sphere is the cornea and a posterior sphere representing the sclera. In order to implement this loss we first estimate a depth map, and then using the semantic segmentation we calculate the sphericity of the two regions: cornea, sclera.

Figure 5.9 – Human eye anatomy [202].



To simplify the calculations we have a sphere fitting implemented for each region. We also apply a threshold before calculating this loss, by counting the number of pixels present for each region and ensuring there is at least $> 50\%$. Although we previously applied a threshold of 40% ROI for all frames, as we mentioned in 5.2, this threshold is to make sure we have enough of each region visible before penalizing its sphericity. The frame is discarded for the sphere fitting loss calculation if this threshold is not met for either regions. We define our threshold as the count of non zero pixels pertaining to either regions and dividing that by the total number of pixels of the frame, Equation 5.27.

Figure 5.10 – Modelling of the eye as two intersecting spheres.



$$m_{ROI} = \begin{cases} m_{cornea} & m_{ROI}(x, y) = 1 \\ m_{sclera} & m_{ROI}(x, y) = 2 \end{cases} \quad (5.26)$$

$$\text{Region percentage } r_i = \frac{\sum_{\forall(x,y) \in m_{ROI}(x,y)} \mathbb{1}_{[m_{ROI}(x,y)=i]}}{\sum_{\forall(x,y) \in m_{ROI}(x,y)} 1} \times 100 \quad (5.27)$$

where $i = 1(\text{cornea})$ or $i = 2(\text{sclera})$

Once the region percentage r_i for regions: cornea, sclera is $> 50\%$ we are able to use the respective frame to calculate the \mathcal{L}_{SFL} . Taking the predicted depth map \hat{D} and the mask ROI m_{ROI} we only keep the depth estimations for that region :

$$\begin{aligned} \hat{D}_{cornea} &= m_{cornea} \times \hat{D} \\ \hat{D}_{sclera} &= m_{sclera} \times \hat{D} \end{aligned} \quad (5.28)$$

Using this regional depth map, and the associated frame we are able to obtain a 3-D point cloud using the inverse of the intrinsic matrix. This geometric projection can be explained by first looking at the simple conversion of world coordinates to image plane coordinates written with homogeneous coordinates. Every point (x, y) in a 2D Cartesian plane has a corresponding set of homogeneous coordinates in the 3D projective space. Once we obtain the pixel co-ordinates as homogeneous coordinates while keeping the last dimension = 1, we can perform any operation or transformation [203]. Equation 5.29 shows the conversion from image plane (pixels), as homogeneous coordinates to world co-ordinates $(x, y, z, 1)$, where $\frac{1}{z}$ is proportional to disparity. Disparity is the difference in horizontal position of a point's projections between a left and right image, Eq ??.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{z} \begin{bmatrix} K & 0 \\ 0 & 1 \end{bmatrix} [R|t] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (5.29)$$

$$\begin{bmatrix} u \\ v \\ 1 \\ \frac{1}{z} \end{bmatrix} = \frac{1}{z} \begin{bmatrix} K & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (5.30)$$

neglecting R,t the Eq.5.30, the inverse of the camera matrix can be simplified to the following:

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = z \begin{bmatrix} \frac{1}{f_x} & 0 & 0 & 0 \\ 0 & \frac{1}{f_y} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \\ \frac{1}{z} \end{bmatrix} \quad (5.31)$$

We then use these estimated point cloud coordinates and apply a least squares sphere fitting. Following a method by Jekel [204] we are able to determine the best sphere center for the given data points. By rearranging the terms in Eq (5.32) we can express the equation in matrix notation and solve for \vec{c} (5.35). By fitting the data points x_i, y_i, z_i we can solve for the centre coordinates of the sphere x_0, y_0, z_0 and the radius r .

$$(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 = r^2 \quad (5.32)$$

$$x^2 + y^2 + z^2 = 2xx_0 + 2yy_0 + 2zz_0 + r^2 - x_0^2 - y_0^2 - z_0^2 \quad (5.33)$$

$$\vec{f} = A\vec{c} \quad (5.34)$$

$$\vec{f} = \begin{bmatrix} x_i^2 + y_i^2 + z_i^2 \\ x_{i+1}^2 + y_{i+1}^2 + z_{i+1}^2 \\ \vdots \\ x_n^2 + y_n^2 + z_n^2 \end{bmatrix} A = \begin{bmatrix} 2x_i & 2y_i & 2z_i & 1 \\ 2x_{i+1} & 2y_{i+1} & 2z_{i+1} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 2x_n & 2y_n & 2z_n & 1 \end{bmatrix} \vec{c} = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \\ r^2 - x_0^2 - y_0^2 - z_0^2 \end{bmatrix} \quad (5.35)$$

$$\vec{f} = \begin{bmatrix} x_k^2 + y_k^2 + z_k^2 \\ x_{k+1}^2 + y_{k+1}^2 + z_{k+1}^2 \\ \vdots \\ x_n^2 + y_n^2 + z_n^2 \end{bmatrix} A = \begin{bmatrix} 2x_k & 2y_k & 2z_k & 1 \\ 2x_{k+1} & 2y_{k+1} & 2z_{k+1} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 2x_n & 2y_n & 2z_n & 1 \end{bmatrix} \vec{c} = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \\ r^2 - x_0^2 - y_0^2 - z_0^2 \end{bmatrix} \quad (5.36)$$

Sphere fitting loss \mathcal{L}_{SFL} is a mean square error (MSE) between the fitted sphere and the data points. The sphericity for each of the corneal and scleral regions have a weight α_e . With both regions fitted to a sphere we then calculate the loss for each pixel p . The threshold applied before calculating this loss ensures either regions are present in the frame. The final loss is a sum of the sphericity of both regions.

$$\mathcal{L}_{SFL} = \mathcal{L}_{cornea_{SFL}} + \mathcal{L}_{sclera_{SFL}} \quad (5.37)$$

$$\mathcal{L}_{cornea_{SFL}} = \frac{1}{p_c} \sum_{k=1}^{p_c} ((x_{ck} - x_{c0}) - r_c)^2, \mathcal{L}_{sclera_{SFL}} = \frac{1}{p_s} \sum_{k=1}^{p_s} ((x_{sk} - x_{s0}) - r_s)^2 \quad (5.38)$$

where p_c are number of pixels, x_{ck} data points, x_{c0} centre coordinate on the corneal surface, r_c the cornea radius and p_s are number of pixels, x_{sk} data points, x_{s0} centre coordinate on the scleral surface, r_s the estimated sclera radius.

Our proposed method now has the following total loss:

$$\mathcal{L}_{total} = \alpha_a \mathcal{L}_{SRL} + \alpha_b \mathcal{L}_{recon} + \alpha_c \mathcal{L}_{SSIM} + \alpha_d \mathcal{L}_{DS} + \alpha_e \mathcal{L}_{SFL} \quad (5.39)$$

where $\alpha_a = 0.85$, $\alpha_b = 0.15$, $\alpha_c = 0.04$., and $\alpha_d = 1$, and $\alpha_e = 10k$.

5.5.2 Training

Once we established this loss we first acknowledged that the calculated errors of sphericity on both regions were very minor. This required a large weight so we could be sure that the loss was making an impact in training. This was done through various tests of changing the α for \mathcal{L}_{SFL} until we saw a stable convergence in the loss. We focused on the total sphere loss, since the loss per region was sometimes ≈ 0 for certain frames. We also kept the frame step of $n = 10$, and the pre-trained model on dataset 'O' fixed during these experiments.

5.5.3 Results

The results are shown in Table 5.12 where we started with the baseline being our best result from the previous section. We then added our novel sphere fitting loss that gave us the

best improvement yet with a difference of $\approx 7.21px, 0.45\%$ between Exp. No. C2 and with an Euclidean of $\approx 12.92px, 0.81\%$ and D4 of $\approx 4.77px, 0.29\%$

Both these results are consistent showing that the addition of this new depth constraint to the training improved results. Most importantly the sphericity constraint seemed to have a bigger impact than the semantic reconstruction loss. Ultimately, this allowed us to obtain results as good as, and even a little better, than the grader errors with a difference of $\approx 0.04px, 0.01\%$ Euclidean distance. By plotting the point cloud we are able to visualise if the constraint had enabled the depth maps to improve. We believe that the results in Section 5.4.2 Table 5.4.2, were enhanced through depth smoothness, alongside the semantic reconstruction loss \mathcal{L}_{SRL} but visually do not correspond perfectly to the shape of the eye, but are a great improvement. Figure 6.2 shows point cloud plot examples from two experiments.

Figure 5.11 – Point cloud plots.

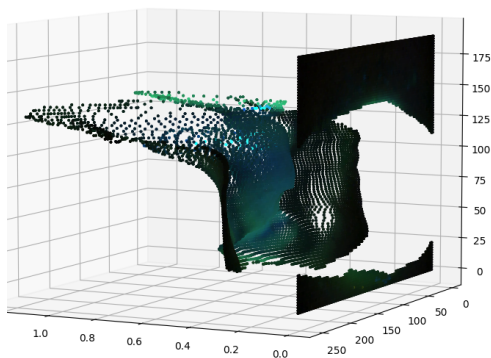
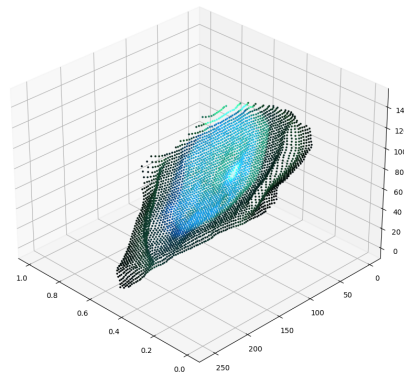
(a) Point without \mathcal{L}_{SFL} , Exp. No. C2.(b) Point cloud using \mathcal{L}_{SFL} , Exp. No. D4.

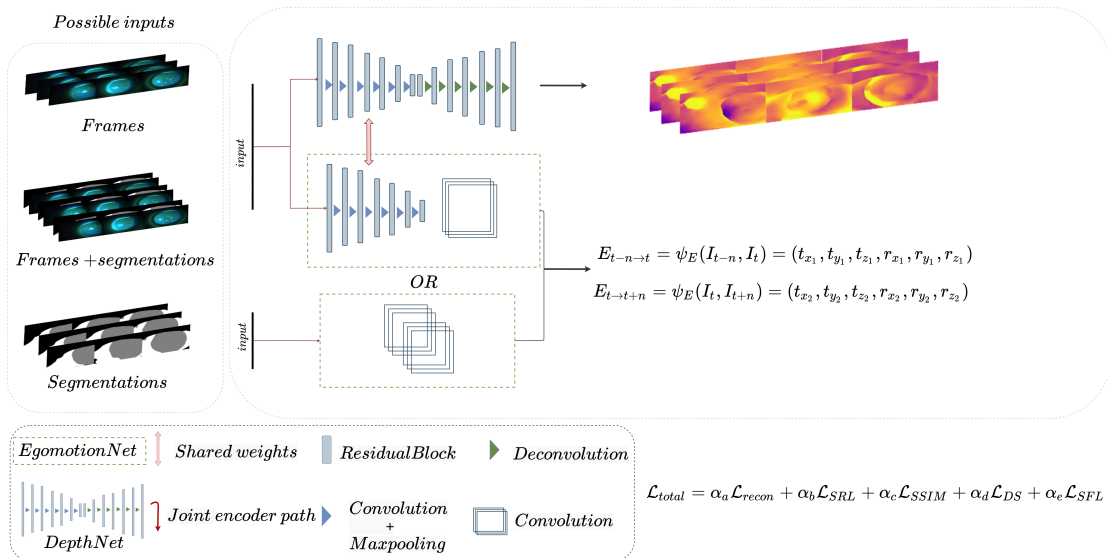
Table 5.12 – Experiment details and results D.

No.	Frame step	DepthNet Input	EgoNet Input	\mathcal{L}_{SRL}	\mathcal{L}_{SFL}	\mathcal{L}_{Recon}	Mean Euclidian (px)	Mean Euclidian (%)	Training dataset	Pre-train dataset
D1	n=10	Frames Binary	Frames Binary	Yes	No	Yes	12.92	0.81	P	O
D2	n=10	Frames Binary	Frames Binary	Yes	Yes	Yes	5.71	0.36	P	O
D3	n=10	Frames Binary	Segmentation	Yes	No	Yes	17.64	1.10	P	O
Grader errors	-	-	-	-	-	-	5.30	0.33	-	-
D4	n=10	Frames Binary	Segmentation	Yes	Yes	Yes	4.77	0.29	P	O

5.6 Conclusion

Our two main contributions have enabled us to obtain a fully self-supervised method we named 'SiGMoid: Semantic & geometric monocular visual odometry'. SiGMoid learns both depth and egomotion, which allows for a successful image registration. Our algorithm with the novel sphere fitting loss demonstrated results not only superior to baselines but also to human error. We implemented sphere shape fitting and did not consider the scale, which can also be included as an additional constraint. The human eye as we mentioned is estimated to have a radius of $\approx 12mm$. By including the scale on the point cloud estimation we can measure how realistic the predictions are. This is a possible improvement on shape fitting that is made possible because of the use of prior knowledge. Sphere fitting loss and the semantic reconstruction loss improved our registration results considerably. We can envision that including the scale can produce a full up to scale 3D reconstruction of the eye. Figure 5.12 illustrates the SiGMoid framework and complete losses. Both CNNs although trained jointly can be used for inference separately. With the help of the manual annotations our evaluation method relies on the inverse warping which validates both the predictions. By finalising a state-of-the-art method to register the frames as well as select viable frames in the pre-processing we move on to applying this to its application in the DED Oxford grade prediction.

Figure 5.12 – SiGMoid: Semantic & geometric monocular visual odometry framework.



OXFORD GRADE CLASSIFICATION USING SIGMOID

“The key to artificial intelligence has always
been the representation.”

— *Jeff Hawkins*

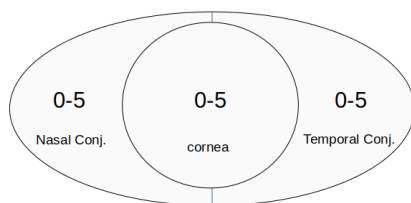
6.1	Introduction	102
6.1.1	Pre-processing	103
6.2	Evaluation metrics	104
6.3	Multi-class Classification	104
6.3.1	Classification Methods	104
6.3.2	Experiments & results	105
6.4	Binary classification	106
6.4.1	Experiments & Results	107
6.5	Conclusion	113

This chapter explores the final task of DED grading prediction. A part of the work was implemented during an internship I supervised. The classification task focused on the Oxford score and we experimented with multi-class and binary classification. Ultimately, we utilise the classification framework to further evaluate SiGMoid, our proposed method.

6.1 Introduction

With a finalised self-supervised algorithm, SiGMoid, we are able to register frames to a common coordinate system. A precise registration can avert over or under estimation of punctate dots grading. We relied on the grading scale, Oxford score, detailed in Chapter 1 to train a classifier to predict the scores. The grading scale annotations were given for the cornea, nasal and temporal conjunctiva with each having a score of 0-5 and the total Oxford score, the sum, can be 0-15 as shown in Figure 6.1. Given our completed semantic segmentation task, detailed in section 5.2, we are able to distinguish these three parts using the 'Mask ROI' m_{ROI} .

Figure 6.1 – Oxford score [54].



We set up some of the following classification tasks as part of an internship that I had the pleasure of supervising. The work that of this internship, by the intern Abdel OUEDRAOGO, is detailed in sections 6.1.1, 6.3.1, and lastly 6.3 which was also further developed later on. We first look over the database 'P' to evaluate the distribution of the videos and their Oxford score. With a total of 16,800 frames we noticed that the grade 5 was the lowest available data example. This indicated that our database is unbalanced for a multi-class classification problem. The automation of grading DED can be accomplished through a classical multi-class classification. We examined various ways to optimize the classification task before incorporating our self-supervised algorithm. Incorporating SiGMoid would enable us to validate that registered frames facilitate the classification task.

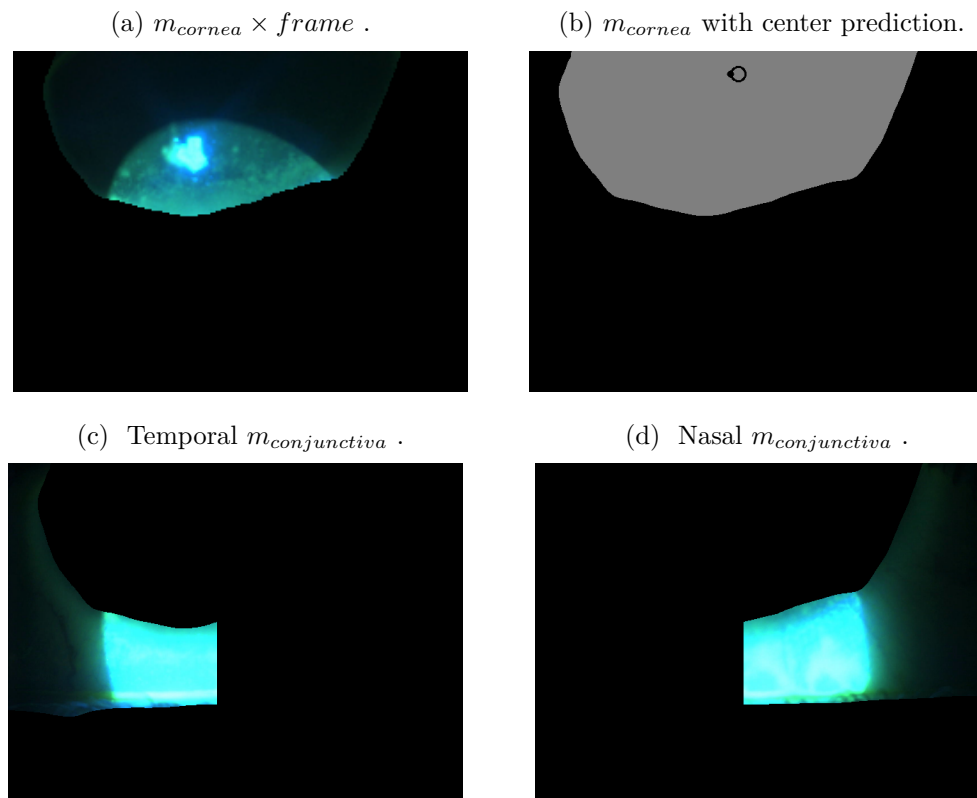
Table 6.1 – Database Oxford score grading.

Oxford score	Videos	Frames	Percentage
0	15	3000	17.1%
1	21	4200	25%
2	14	2800	16.7%
3	21	4200	25%
4	10	2000	11.9%
5	3	600	3.6%

6.1.1 Pre-processing

In order to follow the Oxford grading procedure we first implemented a pre-processing to distinguish the nasal and temporal conjunctiva. Although our m_{ROI} gave us the information of which areas of the frames belonged to the cornea and conjunctiva, we needed an additional method to distinguish between the nasal and temporal conjunctiva. The videos were labelled with a right or left eye label. This facilitated the task the only supplemental information required was to detect the center of the cornea. With this we know that for left eye videos anything left to the cornea is the nasal conjunctiva and to the right was the temporal conjunctiva, and vice versa. A well known algorithm, Kalman filtering [205], predicts the position of a tracked object and updates the series of measurements over a period of time. By extracting only the cornea from m_{cornea} , we predict the center for each and apply it to the frames. The example below is for a left eye video where the center, in pixels, was estimated to be $x = 735, y = 81.25$.

Figure 6.2 – Kalman filtering with left eye video [205].



6.2 Evaluation metrics

Accuracy (ACC), precision and recall were used as metrics for all classification results with the Top-N (N=3) accuracy added for multi-class classification, and Area under the (receiver operating characteristic) curve (AUC) for the binary classification.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6.1)$$

$$Precision = \frac{TP}{TP + FP} \quad (6.2)$$

$$Recall = \frac{TP}{TP + FN} \quad (6.3)$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \quad (6.4)$$

where true positive (TP) is when the prediction is the positive class, true negative (TN) when the prediction is the correct negative class, false positive (FP) the correct class is predicted incorrectly, false negative (FN) where the negative class is predicted incorrectly.

Top-1 accuracy will consider a prediction as correct if and only if the most probable prediction is the correct. Top-N Accuracy takes N predictions with highest probability, we calculated the accuracy for 3 highest probabilities for multi-class prediction.

$$\text{True Positive rate (TPR)} = \frac{TP}{TP + FN} \quad (6.5)$$

$$\text{False Positive rate (FPR)} = \frac{FP}{FP + TN} \quad (6.6)$$

The ROC curve plots TPR or recall vs. FPR, and AUC is the measure of the area under that curve. It measures the performance of various classification thresholds. It also helps interpret the probability that the trained model predicts positives more randomly than negatives. It ranges from 0-1, where 0 is for a model whose predictions are all wrong and 1 are all correct. Both the validation and test evaluation were performed per eye by using a majority vote for the f frames per patient.

6.3 Multi-class Classification

6.3.1 Classification Methods

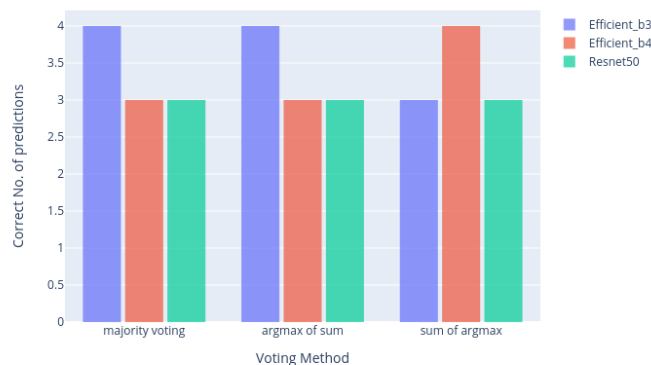
We tested adding a voting method to compare the classification given our limited dataset. The three different methods were the following:

1. Majority voting:
 - (a) Classify each of the frames of all the video.
 - (b) Choose the class to which the majority of the frames belong to.
2. Argmax of sum:
 - (a) Mean of the probability vectors.
 - (b) Choose the class corresponding to the highest probability.
3. Sum of argmax:
 - (a) First group the frames by class.
 - (b) In each class, average the probabilities associated with each frame.
 - (c) Choose the class corresponding to the highest value.

6.3.2 Experiments & results

We tested three commonly used backbones to simply evaluate the voting methods: ResNet50, EfficientNetB3, and EfficientNetB4. Figure 6.3 presents these results, but Tables 6.2,6.3 are more detailed evaluations of the validation and test set. With this experiment we saw the best results using the majority voting method. We decided to use this for all upcoming experiments of classification.

Figure 6.3 – Classical multi-class classification voting method evaluation per patient.



With the great difference in performance between the two sets, we concluded that the multi-class classification was difficult. Our unbalanced dataset was an indication and the results clearly demonstrated an overfitting issue. Overfitting is a common phenomenon where the trained model

Table 6.2 – Validation set classical multi-class classification results.

<i>Backbone</i>	<i>Top-1 Acc</i>	<i>Top-3 Acc</i>	<i>Top-3 Acc</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
ResNet50	0.93	0.99	0.99	0.31	0.27	0.29
EfficientNetB3	0.87	0.99	0.99	0.27	0.24	0.25
EfficientNetB4	0.60	0.81	0.97	0.25	0.12	0.13

Table 6.3 – Test set classical multi-class classification results.

<i>Backbone</i>	<i>Top-1 Acc</i>	<i>Top-3 Acc</i>	<i>Top-3 Acc</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
ResNet50	0.25	0.53	0.78	0.13	0.27	0.15
EfficientNetB3	0.32	0.47	0.54	0.25	0.26	0.23
EfficientNetB4	0.23	0.41	0.55	0.27	0.25	0.15

fits on the training set and is unable to generalize with unseen data. Some explanations include training the model for a long period of time or with few examples causes it to focus on irrelevant information. Once the model is trained, it unfortunately has memorized the training set in the wrong way and cannot perform the prediction or classification it was trained for effectively.

Given distribution of our complete database, shown in table 6.1, and following the poor results obtained for multi-class classification, we decided to focus on a binary classification (mild DED vs. severe DED) to evaluate our proposed method in a diagnostic aspect.

6.4 Binary classification

Now the model is trained to classify whether the patient has a mild DED or a severe DED. Mild DED (respectively severe DED) is defined as a corneal Oxford score ≤ 1 (respectively ≥ 2) [206]. We also decided to focus on the cornea grade classification to follow the Oxford score grading technique. As was done for the training, which used all frames from each patient’s video, we continued to validate and test per patient. Once the model is trained, it predicts scores for a set of frames and the majority vote is then applied to obtain the final predicted grade.

For the following classification evaluation we first had to outline a data split that would respect several rules. By training with only database ‘P’ we found that there wasn’t enough frames per class or even overall. Again we decided to utilise the original database ‘O’ to help in balancing and also take advantage of extra samples to train with. Given that database ‘P’ is our primary training set we first formed a data split where the database ‘P’ was the primary training set, keeping the patients originally placed and used as test for the SiGMoid implementation in the test set. We then placed all the data from database ‘O’, containing 32 patients, as the validation set, detailed in Table 6.4.

An additional detail in the tables is the pre-processing: in one scenario, named ‘Select best’, frame selection is performed, using SiGMoid. This was realized using the warping error which the semantic registration loss \mathcal{L}_{SRL} we detailed in Chapter 5. Once the warping error is calculated for all the dataset, we set a threshold to only keep frames with a *WarpingError* $< 5\%$. An average of 35% of frames were removed once this limitation was applied.

Table 6.4 – Data split description.

Set	No. Patients	No. of eyes	Frames	Percentage	Best Frames	Best Percentage
Train	18	35	9,998	38.1%	6,425	37.3%
Validation	32	41	11,336	43.2%	7,327	42.5%
Test	7	14	4,911	18.7%	3,487	20.2%

Once we established the data splits the experiments were performed using two modes. Classical classification ‘C’ where the input was the frames of the cornea $frames \times m_{cornea}$. SiGMoid classification ‘SG’ where the input was the fusion of two frames giving us a mosaic from the pair of frames. An example of the training frames used as an input for both modes is shown below in Figures 6.5, 6.6.

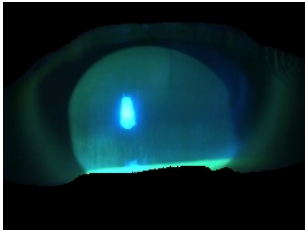
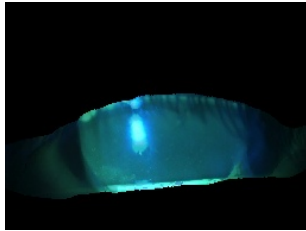
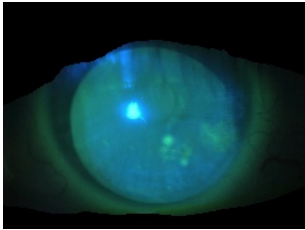
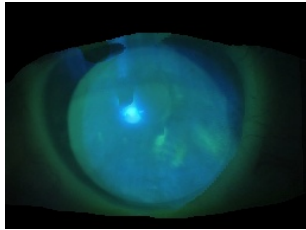
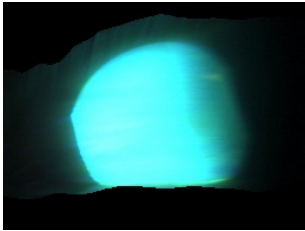
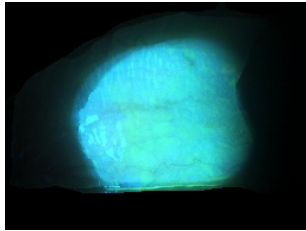
6.4.1 Experiments & Results

For the experiments we attempted to investigate various hyper-parameters including :

- Backbone including : Resnet50, Densenet121, Inception V3, NoisyStudent (EfficientNet-B4)
- Learning rate: $[2e^{-3} - 2e^{-6}]$
- LR schedulers : OneCycleLR, StepLR, CosineAnnealingLR
- Weight decay: $[0.1 - 3e^{-4}]$
- Select best (True/False)
- Div factor (DF), Step Epoch (SE)

These are a few of many hyper-parameters that can be explored when fine-tuning. We deemed these the most important and, solely due to time constraints, we based our experiments on changing them. To better clarify the backbone of the experiments includes existing architectures. The learning rate (LR) is a key parameter that determines the step size and manages how quickly the model can adapt to the problem. Weight decay is a regularization technique to better moderate how the weights of the CNN are obtained.

Table 6.5 – Classification input examples.

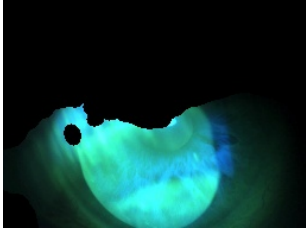
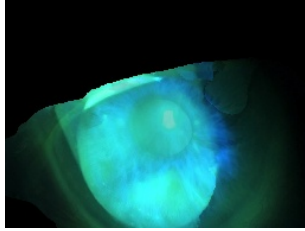
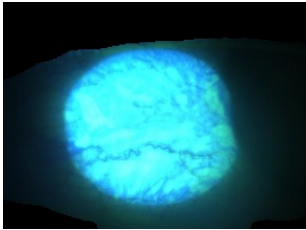
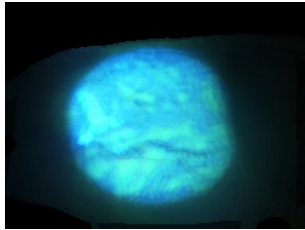
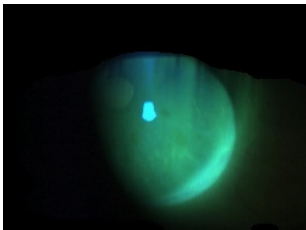
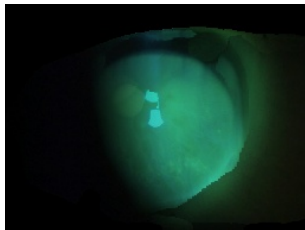
Classical input	SiGMoid input
	
	
	

We tested OneCycleLR which is an optimizer that changes the LR after every batch. This policy was described in [207] and it modifies the LR from an initial value to a maximum, which can be the initial LR set at the beginning of training, and then to a minimum value. This is done based on the “Div factor” which we also included as a hyper-parameter. This is determined by the following Equation 6.7. This policy allows for an online learning rate optimization.

$$\text{Initial LR} = \frac{\text{Maximum LR}}{\text{Div factor}} \quad (6.7)$$

Another optimizer is the StepLR that decays the learning rate by gamma every step epoch. We maintained the gamma at = 0.1. Lastly, the CosineAnnealingLR proposed in [208] that begins with a large LR and aggressively decreases it before increasing it. It also maintains information from the previous cycle each time it restarts, the Equation 6.8 below details the new LR which also relies on the step epoch.

Table 6.6 – Classification input examples II .

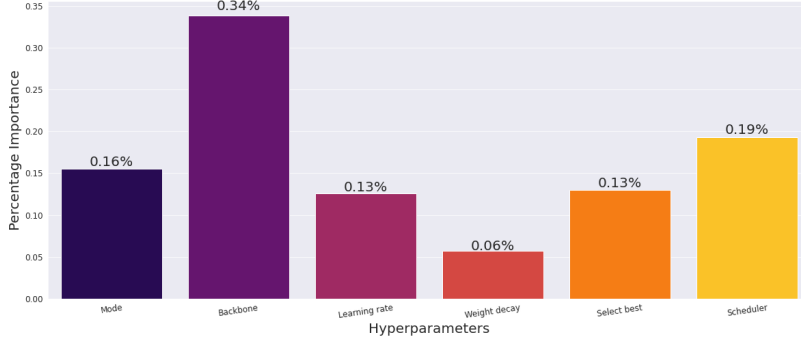
Classical input	SiGMoid input
	
	
	

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min}) \left(1 + \cos \left(\frac{T_{cur}}{T_{max}} \pi \right) \right) \quad (6.8)$$

We implemented for each of the backbones, while varying each of these hyper-parameters. This amounted to 179 results. For the full set of experiments we calculated the importance of each of the hyperparameters. Figure 6.4 displays the percentage importance of each of these parameters. We note that the most prominent being the backbone but that the scheduler and the mode are also important and the fourth being the ‘Select best’. This shows that the Mode: ‘SG’ classification using frames registered using SiGMoid or ‘C’ classical classification using raw frames as input, along with ‘Select best’ have an impact on the classification results. It also validates the need for such a diverse set of experiments with different backbones and schedulers.

The best results for both methods and their respected configurations are shown below in Table 6.7. For mode SG, an AUC of 0.93 was the highest; an AUC of 0.64 was obtained for mode C with the same configuration. For mode C, the highest AUC obtained was 0.87; an AUC

Figure 6.4 – Hyperparameters percentage importance for the classification results.



of 0.67 was obtained for mode SG with the same configuration, but the accuracy, precision, or recall remained constant.

Table 6.7 – Best configuration test set results.

Mode	Backbone	LR	Weight decay	Select best	DF	Scheduler	SE	AUC	Acc	Precision	Recall
SG	NoisyStudent (EfficientNet-B4)	0.0002	0.0003	True	NA	CosineAnnealingLR	15	0.93	0.64	0.75	0.72
C	NoisyStudent (EfficientNet-B4)	0.0002	0.0003	False	NA	CosineAnnealingLR	15	0.64	0.43	0.70	0.56
C	Inception V3	0.002	0.0003	False	20.0	OneCycleLR	NA	0.87	0.36	0.18	0.50
SG	Inception V3	0.002	0.0003	False	20.0	OneCycleLR	NA	0.67	0.36	0.18	0.50

We display the detailed results tables in the appendix but plotted them for easier interpretation and for discussion purposes. Our first scatter plots display two sets of results, and for the experiment number (Exp. No.) the first 40 are using the SiGMoid input and the remaining 40-80 are using the classical frame input. We also added a boxplot to summarise the median, min and max obtained by both modes, SG and C, for each backbone. For the setup where both methods had ‘select best = False’; Figures 6.5, 6.6 present the results and Figures 6.7, 6.8 for when we set ‘select best = True’ for the SiGMoid input classification.

For the first set of experiments, ‘select best = False’, the highest AUC was obtained with mode SG and an overall trend of the SG experiments can be seen to maintain an average AUC above 0.5 in Figure 6.5. There are also more experiments under an AUC of 0.3 with the mode C. As for the box plots the medians are close for two backbones: Densenet121 and NoisyStudent (EfficientNet-B4), but higher with mode SG for Resnet50, and Inception V3, shown in Figure 6.6.

We then incorporated, ‘select best = True’ for the SiGMoid classification as it helps us reduce frames that were deemed to have been badly registered. As we mentioned, this was based on a warping error. With less data for the SiGMoid training we were still able to obtain four of the highest AUCs and also maintain an average that was higher than that of the classical classification mode C. The overall distribution for mode SG in Figure 6.7 is more consistent

Figure 6.5 – Scatter plot with both modes having select best = False.



Figure 6.6 – Boxplot with both modes having select best = False.

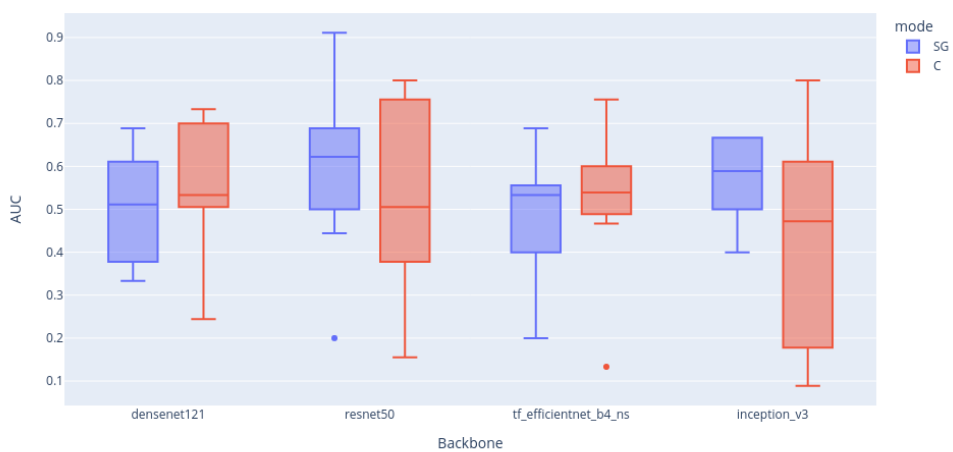
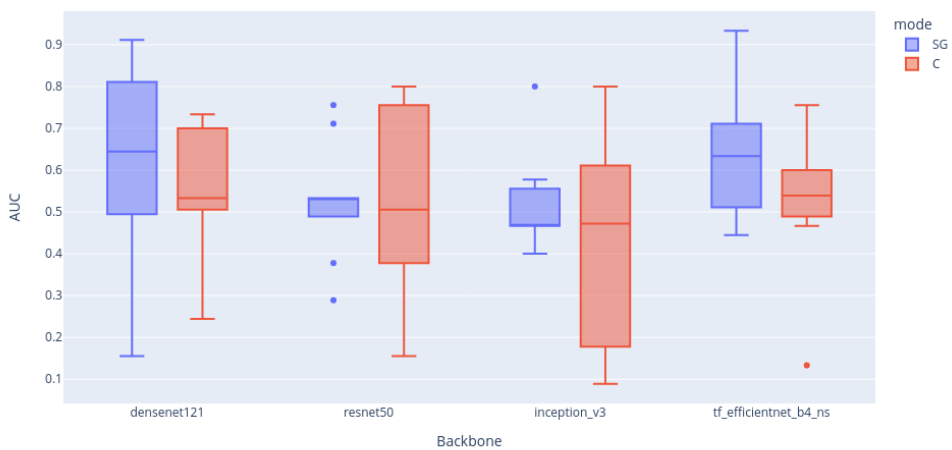


Figure 6.7 – Scatter plot with SiGMoid select best = True.



Figure 6.8 – Boxplot with only SiGMoid select best = True.



and we also observe more experiments in mode C that are below an AUC of 0.3. The boxplots in Figure 6.8 show that the SG method is enhanced with addition of frame selection. With a superior median for the backbones: Densenet121 and NoisyStudent (EfficientNet-B4) which were only close in the first experiment. As for Resnet50 and Inception V3 experiments, the medians were close with both modes.

6.5 Conclusion

In the previous chapter, we presented SiGMoid, a self-supervised image registration algorithm towards DED diagnosis and quantification from slit lamp videos. In this chapter, we demonstrated that obtaining an accurate reconstruction is beneficial to the classification of DED grading. It helps in two ways by enlarging the field of view via the registration and allows removal of poor quality frames. Both advantages put together lead to an improvement in classification performance. We demonstrated the generalizability of our method's classification improvement through a large number of experiments.

CONCLUSION & PERSPECTIVES

“If we knew what we were doing, it would not be called research, would it?”

— *Albert Einstein*

Conclusion

Sjogren’s syndrome (SS) is a disorder of the immune system that critically requires more optimized and precise grading of its symptoms. Two of the main symptoms being dry mouth and dry eye. In this work we focused on Dry eye disease (DED), which is also a disease that can be present outside of SS with a prevalence increasing with age. Found to be higher in women than men, prevalence is found to have no correlation with education or location [209]. A comprehensive study on existing methods that we published allowed us to better comprehend what has been done in the field and what was missing. Most existing diagnostic methods are unfortunately invasive, non-reproducible, lacking in accuracy and constantly subjective. Methods are expected to aid in evaluating the progress of the patient, the extent of the damage, and how well certain medications perform. Only a handful of recent methods are able to automate quantification and obtain a grade without clinician interference. We found that although evolution towards automated quantification of DED is slow, promising results have been obtained. The next step seems to be the integration of artificial intelligence (AI).

The focal point of our work was to obtain a method that can be deployed with minimal additional information and aid in DED diagnosis. Our main proposition was to utilise the videos obtained from DED staining examinations and try to improve the visualisation of the damaged areas. We believe that this could also help in avoiding over or under-estimation when grading. With this we prioritize the use of projective geometry and visual odometry. A field that exploited this area significantly is autonomous driving and robotics. We therefore explored the domain, and narrowed our interest to methods that can be implemented in an unsupervised or self-supervised manner. Along the investigation we performed various preliminary tasks before finalising our proposed method SiGMoid, and lastly applied our method to DED grade prediction for evaluation.

With two main databases obtained we investigated methods that learn to register images to a common coordinate system. These methods focused on learning and estimating depth from

images and camera motion between pairs of images. While keeping in mind the main limitations we have with our examination videos. These challenges included (1) the heterogeneity of our data, (2) disturbances in the data due to specular reflections. We introduced existing methods in Chapter 2, some of which we focused on and re-implemented in Chapter 5. The methods we focused on had a common supervision signal based on depth image based rendering, namely the photometric loss. With this we were able to validate that the disturbances present in our data hindered training. The main complication caused training to diverge to a loss of infinity or zero.

To address these complications we integrated semantic segmentation masks in two different ways. We first introduced masks that alienated any disturbances and unwanted information, allowing the models to only focus on well lit areas. Our second masks highlighted the eye and concealed the eyelid. We then tested the baseline with both masks and were able to obtain a stable training, but still an unsuccessful one. With no sign that the total loss was improving or converging we then focused on the main supervision signals and the assumptions made when they are used. We found that we violated most assumptions leading us to our first contribution, the semantic reconstruction loss. We replaced the photometric loss as the main loss, with the semantic reconstruction loss seeing that it allowed us to overlook the major disturbances present in our data when learning the 3D camera motion (egomotion). With baseline results we were able to see improvement using manual annotations of punctate dots (damaged eye areas) and veins. Once we predicted a transformation between frames we were able to measure the error in Euclidean distance (pixels, percentage).

Once we had a constraint that was more suitable to our data for the egomotion, we focused on the depth estimation. With this we introduced our more valuable contribution which was derived from the notion that we have greater knowledge about the object seen in our videos unlike in autonomous driving. We then considered how we may use the eye’s anatomy as a restriction for the depth model. Therefore, we developed a loss that penalizes the estimated point cloud, obtained from the depth estimations, to fit a more spherical shape. Since the cornea and conjunctiva are two intersecting spheres that make up the human eye, we implemented a sphericity loss that penalized each. With this final incorporation our proposed method was finalised, which we named ‘SiGMoid: Semantic & geometric monocular visual odometry’. We were also able to patent the key idea behind the shape fitting loss as a ‘Method for modeling an anatomical structure of the human body’. Furthermore, once this was evaluated we obtained more noticeable improvements. Considering that we used a human error as a benchmark, which we obtained via interobserver variability of the annotations used, SiGMoid obtained marginally superior results.

Lastly, we proposed an application to DED grading through classification of the Oxford score [54]. In this work the first tasks were carried out as a part of an internship that was completed

by the intern: Abdel OUEDRAOGO. We first investigated classical classification where frames from the examinations were used to predict the Oxford score. We also looked at various classification methods, as well as experimented with different backbones and hyper-parameters. With a challenging and unbalanced dataset we focused on binary classification. We then extended Abdel's work and performed a more thorough analysis to evaluate the incorporation of SiGMoid to the classification. We saw improvement on with various backbones and hyper-parameters which validated the generalizability. Ultimately we were able to show that the classification of DED grading benefits from having an accurate reconstruction.

Perspectives

We believe that it is crucial to increase both the visual quality and visual field in order to have a basis for better DED grading. By applying deep learning methods that are less conventional we were able to find a self-supervised method to register frames to a common coordinate system. DED has yet to benefit from all types of advances we have seen grow over the years in AI. We think that this thesis has shown, through encouraging findings and original concepts, that there is still more to learn about this topic because it is slowly growing with AI infused interest.

Some of the perspectives we wish to address in the future:

1. Recently, our team was granted access to a more modern database. This database was obtained utilizing a more sophisticated acquisition protocol, and the quality is evidently much better. With this, we would be able to train with examinations of better quality in addition to expanding our training set. From Qauntel medical, the ION imaging system¹ produces high resolution videos and images (4K). The ION system combines the slit-lamp with Apple's iPhone technology allowing for fully connected examinations.
2. We believe that the sphericity constraint can be expanded by including the scale. Once we input the scale we can further utilise the human eye estimate radius to also help the model predict more realistic depth maps. This can also ultimately lead to more accurate 3D reconstructions.
3. Another way to expand the sphericity loss is through a sphericity measure ψ constraint [210]. Defined as the ratio of the surface area and volume, it is 1 for a sphere and less for any other shape. A study in [211] focused on both sphericity and roundness which can also be incorporated to as a differentiation constraint.
4. Apply our proposed solutions to other DED diagnostic methods for a richer analysis. This is possible as we have annotations for various grades as we mentioned in Chapter 3,

1. <https://www.quantel-medical.fr/products/ophthalmology/diagnostique/ion>

including TBUT, Schirmer, and OSS. These can also be integrated for a Multi-modality approach. Along with the inclusion of demographic data we can obtain more reliable predictions.

5. Additionally having patient follow-up examinations could allow us to expand our longitudinal data. With this we can ensure the detection of even subtle changes, or response to treatment.

In future studies we would aim to replicate results with a larger database. The possibility of obtaining a new grading scale warrants further investigation. We have demonstrated the great potential of this field, and our findings have been evaluated and found to support our theories.

PUBLICATIONS

Below are the list of publications completed during this PhD:

Journal publications:

- Brahim, I., Lamard, M., Benyoussef, A. A., & Quellec, G. (2022). Automation of dry eye disease quantitative assessment: A review. *Clinical & Experimental Ophthalmology*.

Conference papers:

- Brahim, I., Lamard, M., Benyoussef, A. A., Conze, P. H., Cochener, B., Cornec, D., & Quellec, G. (2022). Mapping the ocular surface from monocular videos with an application to dry eye disease grading. In *International Workshop on Ophthalmic Medical Image Analysis* (pp. 63-72). Springer, Cham.
- Brahim, I., Lamard, M., Benyoussef, A.-A., Conze, P.-H., Cochener, B., Quellec, G., & Cornec, D. (2022) “15th International Symposium on Sjögren’s syndrome” *Clinical and Experimental Rheumatology*.

Patents:

- Brahim, I., Lamard, M., & Quellec, G. (2022). Method for modeling an anatomical structure of the human body. Brevet européen réf: DV 4731 (réf: BR 130873/BAN/MIC/EVG)

APPENDIX

Table 6.8 – Detailed classification Test set experiment results using S1 data split - Part I.

Mode	Backbone	LR	Weight decay	Select Best	DF	Scheduler	SE	AUC	Acc	Precision	Recall
SG	densenet121	0.0002	0.0003	True	15.0	OneCycleLR	NA	0.91	0.57	0.73	0.67
SG	densenet121	0.0002	0.0003	False	15.0	OneCycleLR	NA	0.51	0.5	0.57	0.57
C	densenet121	0.0002	0.0003	False	15.0	OneCycleLR	NA	0.73	0.64	0.75	0.72
SG	densenet121	0.002	0.0003	True	20.0	OneCycleLR	NA	0.47	0.36	0.4	0.41
SG	densenet121	0.002	0.0003	False	20.0	OneCycleLR	NA	0.64	0.36	0.18	0.5
C	densenet121	0.002	0.0003	False	20.0	OneCycleLR	NA	0.51	0.36	0.18	0.5
SG	densenet121	0.0002	0.0003	True	NA	StepLR	20	0.8	0.36	0.18	0.5
SG	densenet121	0.0002	0.0003	False	NA	StepLR	20	0.49	0.36	0.18	0.5
C	densenet121	0.0002	0.0003	False	NA	StepLR	20	0.73	0.5	0.71	0.61
SG	densenet121	2.00E-06	0.0003	True	NA	StepLR	20	0.64	0.57	0.62	0.62
SG	densenet121	2.00E-06	0.0003	False	NA	StepLR	20	0.69	0.57	0.62	0.62
C	densenet121	2.00E-06	0.0003	False	NA	StepLR	20	0.53	0.64	0.75	0.72
SG	densenet121	0.002	0.001	True	NA	LambdaLR	5	0.5	0.36	0.18	0.5
SG	densenet121	0.002	0.001	False	NA	LambdaLR	5	0.38	0.5	0.29	0.39
C	densenet121	0.002	0.001	False	NA	LambdaLR	5	0.5	0.36	0.18	0.5
SG	densenet121	0.0002	0.0003	True	5.0	OneCycleLR	NA	0.78	0.36	0.18	0.5
SG	densenet121	0.0002	0.0003	False	5.0	OneCycleLR	NA	0.67	0.57	0.73	0.67
C	densenet121	0.0002	0.0003	False	5.0	OneCycleLR	NA	0.67	0.57	0.73	0.67
SG	densenet121	0.0002	0.0003	True	40.0	OneCycleLR	NA	0.71	0.57	0.62	0.62
SG	densenet121	0.0002	0.0003	False	40.0	OneCycleLR	NA	0.33	0.36	0.18	0.5
C	densenet121	0.0002	0.0003	False	40.0	OneCycleLR	NA	0.67	0.36	0.18	0.5
SG	densenet121	0.0002	0.0003	True	NA	CosineAnnealingLR	15	0.82	0.36	0.18	0.5
SG	densenet121	0.0002	0.0003	False	NA	CosineAnnealingLR	15	0.58	0.36	0.18	0.5
C	densenet121	0.0002	0.0003	False	NA	CosineAnnealingLR	15	0.73	0.64	0.75	0.72
SG	densenet121	0.0002	0.001	True	NA	CosineAnnealingLR	15	0.82	0.36	0.18	0.5
SG	densenet121	0.0002	0.001	False	NA	CosineAnnealingLR	15	0.4	0.43	0.52	0.51
C	densenet121	0.0002	0.001	False	NA	CosineAnnealingLR	15	0.73	0.64	0.75	0.72
SG	densenet121	0.0002	0.0003	True	NA	LambdaLR	5	0.53	0.36	0.18	0.5
SG	densenet121	0.0002	0.0003	False	NA	LambdaLR	5	0.38	0.5	0.29	0.39
C	densenet121	0.0002	0.0003	False	NA	LambdaLR	5	0.42	0.36	0.18	0.5
SG	resnet50	0.0002	0.0003	True	15.0	OneCycleLR	NA	0.29	0.21	0.23	0.21
SG	resnet50	0.0002	0.0003	False	15.0	OneCycleLR	NA	0.62	0.64	0.75	0.72
C	resnet50	0.0002	0.0003	False	15.0	OneCycleLR	NA	0.69	0.71	0.78	0.78
SG	resnet50	0.002	0.0003	True	20.0	OneCycleLR	NA	0.53	0.36	0.18	0.5
SG	resnet50	0.002	0.0003	False	20.0	OneCycleLR	NA	0.44	0.57	0.62	0.62
C	resnet50	0.002	0.0003	False	20.0	OneCycleLR	NA	0.47	0.36	0.18	0.5
SG	resnet50	0.0002	0.0003	True	NA	StepLR	20	0.53	0.5	0.57	0.57
SG	resnet50	0.0002	0.0003	False	NA	StepLR	20	0.64	0.57	0.73	0.67
C	resnet50	0.0002	0.0003	False	NA	StepLR	20	0.51	0.64	0.75	0.72
SG	resnet50	2.00E-06	0.0003	True	NA	StepL	20	0.53	0.5	0.71	0.61
SG	resnet50	2.00E-06	0.0003	False	NA	StepLR	20	0.6	0.43	0.52	0.51
C	resnet50	2.00E-06	0.0003	False	NA	StepLR	20	0.16	0.36	0.18	0.5
SG	resnet50	0.002	0.001	True	NA	LambdaLR	5	0.5	0.36	0.18	0.5
SG	resnet50	0.002	0.001	False	NA	LambdaLR	5	0.5	0.36	0.18	0.5
C	resnet50	0.002	0.001	False	NA	LambdaLR	5	0.5	0.36	0.18	0.5
SG	resnet50	0.0002	0.0003	True	5.0	OneCycleLR	NA	0.71	0.57	0.73	0.67
SG	resnet50	0.0002	0.0003	False	5.0	OneCycleLR	NA	0.62	0.64	0.75	0.72
C	resnet50	0.0002	0.0003	False	5.0	OneCycleLR	NA	0.8	0.79	0.81	0.83
SG	resnet50	0.0002	0.0003	True	40.0	OneCycleLR	NA	0.53	0.5	0.57	0.57
SG	resnet50	0.0002	0.0003	False	40.0	OneCycleLR	NA	0.73	0.64	0.75	0.72
C	resnet50	0.0002	0.0003	False	40.0	OneCycleLR	NA	0.76	0.64	0.75	0.72
SG	resnet50	0.0002	0.0003	True	NA	CosineAnnealingLR	15	0.76	0.57	0.73	0.67
SG	resnet50	0.0002	0.0003	False	NA	CosineAnnealingLR	15	0.91	0.64	0.75	0.72
C	resnet50	0.0002	0.0003	False	NA	CosineAnnealingLR	15	0.38	0.36	0.18	0.5
SG	resnet50	0.0002	0.001	True	NA	CosineAnnealingLR	15	0.49	0.5	0.52	0.52
SG	resnet50	0.0002	0.001	False	NA	CosineAnnealingLR	15	0.69	0.57	0.73	0.67
C	resnet50	0.0002	0.001	False	NA	CosineAnnealingLR	15	0.76	0.64	0.75	0.72
SG	resnet50	0.0002	0.0003	True	NA	LambdaLR	5	0.38	0.36	0.18	0.5
SG	resnet50	0.0002	0.0003	False	NA	LambdaLR	5	0.2	0.36	0.18	0.5
C	resnet50	0.0002	0.0003	False	NA	LambdaLR	5	0.18	0.36	0.18	0.5

Table 6.9 – Detailed classification Test set experiment results using S1 data split - Part II.

Mode	Backbone	LR	Weight decay	Select Best	DF	Scheduler	SE	AUC	Acc	Precision	Recall
SG	inception v3	0.0002	0.0003	True	15.0	OneCycleL	NA	0.8	0.57	0.73	0.67
SG	inception v3	0.0002	0.0003	False	15.0	OneCycleL	NA	0.58	0.64	0.75	0.72
C	inception v3	0.0002	0.0003	False	15.0	OneCycleL	NA	0.8	0.71	0.78	0.78
SG	inception v3	0.002	0.0003	True	20.0	OneCycleL	NA	0.58	0.36	0.18	0.5
SG	inception v3	0.002	0.0003	False	20.0	OneCycleL	NA	0.67	0.36	0.18	0.5
C	inception v3	0.002	0.0003	False	20.0	OneCycleL	NA	0.87	0.36	0.18	0.5
SG	inception v3	0.0002	0.0003	True	NA	StepLR	20	0.47	0.36	0.18	0.5
SG	inception v3	0.0002	0.0003	False	NA	StepLR	20	0.67	0.36	0.18	0.5
C	inception v3	0.0002	0.0003	False	NA	StepLR	20	0.18	0.36	0.18	0.5
SG	inception v3	2.00E-06	0.0003	True	NA	StepLR	20	0.56	0.5	0.57	0.57
SG	inception v3	2.00E-06	0.0003	False	NA	StepLR	20	0.6	0.43	0.52	0.51
C	inception v3	2.00E-06	0.0003	False	NA	StepLR	20	0.71	0.5	0.71	0.61
SG	inception v3	0.002	0.001	True	NA	LambdaLR	5	0.5	0.36	0.18	0.5
SG	inception v3	0.002	0.001	False	NA	LambdaLR	5	0.5	0.36	0.18	0.5
C	inception v3	0.002	0.001	False	NA	LambdaLR	5	0.5	0.36	0.18	0.5
SG	inception v3	0.0002	0.0003	True	5.0	OneCycleL	NA	0.47	0.36	0.18	0.5
SG	inception v3	0.0002	0.0003	False	5.0	OneCycleL	NA	0.4	0.36	0.18	0.5
C	inception v3	0.0002	0.0003	False	5.0	OneCycleL	NA	0.51	0.36	0.18	0.5
SG	inception v3	0.0002	0.0003	True	40.0	OneCycleL	NA	0.44	0.5	0.48	0.48
SG	inception v3	0.0002	0.0003	False	40.0	OneCycleL	NA	0.58	0.57	0.73	0.67
C	inception v3	0.0002	0.0003	False	40.0	OneCycleL	NA	0.44	0.5	0.48	0.48
SG	inception v3	0.0002	0.0003	True	NA	CosineAnnealingLR	15	0.47	0.36	0.18	0.5
SG	inception v3	0.0002	0.0003	False	NA	CosineAnnealingLR	15	0.67	0.36	0.18	0.5
C	inception v3	0.0002	0.0003	False	NA	CosineAnnealingLR	15	0.18	0.36	0.18	0.5
SG	inception v3	0.0002	0.001	True	NA	CosineAnnealingLR	15	0.47	0.36	0.18	0.5
SG	inception v3	0.0002	0.001	False	NA	CosineAnnealingLR	15	0.67	0.36	0.18	0.5
C	inception v3	0.0002	0.001	False	NA	CosineAnnealingLR	15	0.16	0.36	0.18	0.5
SG	inception v3	0.0002	0.0003	True	NA	LambdaLR	5	0.4	0.36	0.18	0.5
SG	inception v3	0.0002	0.0003	False	NA	LambdaLR	5	0.43	0.36	0.18	0.5
C	inception v3	0.0002	0.0003	False	NA	LambdaLR	5	0.61	0.36	0.18	0.5
SG	NS Efficientnet b4	0.0002	0.0003	True	15.0	OneCycleL	NA	0.77	0.57	0.73	0.67
SG	NS Efficientnet b4	0.0002	0.0003	False	15.0	OneCycleL	NA	0.67	0.57	0.73	0.67
C	NS Efficientnet b4	0.0002	0.0003	False	15.0	OneCycleL	NA	0.76	0.71	0.7	0.64
SG	NS Efficientnet b4	0.002	0.0003	True	20.0	OneCycleL	NA	0.58	0.5	0.71	0.61
SG	NS Efficientnet b4	0.002	0.0003	False	20.0	OneCycleL	NA	0.4	0.36	0.42	0.46
C	NS Efficientnet b4	0.002	0.0003	False	20.0	OneCycleL	NA	0.13	0.36	0.18	0.5
SG	NS Efficientnet b4	0.0002	0.0003	True	NA	StepLR	20	0.44	0.5	0.71	0.61
SG	NS Efficientnet b4	0.0002	0.0003	False	NA	StepLR	20	0.2	0.29	0.15	0.4
C	NS Efficientnet b4	0.0002	0.0003	False	NA	StepLR	20	0.47	0.36	0.18	0.5
SG	NS Efficientnet b4	2.00E-06	0.0003	True	NA	StepLR	20	0.6	0.29	0.15	0.4
SG	NS Efficientnet b4	2.00E-06	0.0003	False	NA	StepLR	20	0.51	0.36	0.42	0.46
C	NS Efficientnet b4	2.00E-06	0.0003	False	NA	StepLR	20	0.6	0.5	0.57	0.57
SG	NS Efficientnet b4	0.002	0.001	True	NA	LambdaLR	5	0.71	0.64	0.58	0.54
SG	NS Efficientnet b4	0.002	0.001	False	NA	LambdaLR	5	0.56	0.64	0.32	0.5
C	NS Efficientnet b4	0.002	0.001	False	NA	LambdaLR	5	0.53	0.57	0.48	0.49
SG	NS Efficientnet b4	0.0002	0.0003	True	5.0	OneCycleL	NA	0.67	0.5	0.71	0.61
SG	NS Efficientnet b4	0.0002	0.0003	False	5.0	OneCycleL	NA	0.69	0.57	0.73	0.67
C	NS Efficientnet b4	0.0002	0.0003	False	5.0	OneCycleL	NA	0.54	0.43	0.69	0.56
SG	NS Efficientnet b4	0.0002	0.0003	True	40.0	OneCycleL	NA	0.51	0.57	0.62	0.62
SG	NS Efficientnet b4	0.0002	0.0003	False	40.0	OneCycleL	NA	0.56	0.5	0.48	0.48
C	NS Efficientnet b4	0.0002	0.0003	False	40.0	OneCycleL	NA	0.6	0.57	0.57	0.58
SG	NS Efficientnet b4	0.0002	0.0003	True	NA	CosineAnnealingLR	15	0.93	0.64	0.75	0.72
SG	NS Efficientnet b4	0.0002	0.0003	False	NA	CosineAnnealingLR	15	0.44	0.43	0.69	0.56
C	NS Efficientnet b4	0.0002	0.0003	False	NA	CosineAnnealingLR	15	0.64	0.43	0.69	0.56
SG	NS Efficientnet b4	0.0002	0.001	True	NA	CosineAnnealingLR	15	0.51	0.43	0.47	0.47
SG	NS Efficientnet b4	0.0002	0.001	False	NA	CosineAnnealingLR	15	0.29	0.36	0.18	0.5
C	NS Efficientnet b4	0.0002	0.001	False	NA	CosineAnnealingLR	15	0.49	0.36	0.18	0.5
SG	NS Efficientnet b4	0.0002	0.0003	True	NA	LambdaLR	5	0.71	0.64	0.58	0.54
SG	NS Efficientnet b4	0.0002	0.0003	False	NA	LambdaLR	5	0.56	0.64	0.32	0.5
C	NS Efficientnet b4	0.0002	0.0003	False	NA	LambdaLR	5	0.53	0.57	0.48	0.49

BIBLIOGRAPHY

1. Nichols, K. K. *et al.*, Impact of dry eye disease on work productivity, and patients' satisfaction with over-the-counter dry eye treatments, *Investigative ophthalmology & visual science* **57**, 2975–2982 (2016).
2. Craig, J. P. *et al.*, TFOS DEWS II definition and classification report, *The ocular surface* **15**, 276–283 (2017).
3. Uchino, M. & Schaumberg, D. A., Dry eye disease: impact on quality of life and vision, *Current ophthalmology reports* **1**, 51–57 (2013).
4. Brahim, I., Lamard, M., Benyoussef, A.-A. & Quellec, G., Automation of dry eye disease quantitative assessment: A review, *Clinical & Experimental Ophthalmology* (2022).
5. Bron, A., Evans, V. E. & Smith, J. A., Grading of corneal and conjunctival staining in the context of other dry eye tests, *Cornea* **22**, 640–650 (2003).
6. Deng, L., Yu, D., *et al.*, Deep learning: methods and applications, *Foundations and Trends® in Signal Processing* **7**, 197–387 (2014).
7. Ngiam, J. *et al.*, Multimodal deep learning (2011).
8. Litjens, G. *et al.*, A survey on deep learning in medical image analysis, *Medical image analysis* **42**, 60–88 (2017).
9. Lemp, M. A., The Definition Classification of Dry Eye Disease, 6.
10. Wombat, W. & Koala, K., The true meaning of 42, *Journal of modern skepticism* (2016).
11. Hassani, R. T. J., Baudouin, C. & Denoyer, A., *in Surface Oculaire* 139–158 (Elsevier Masson Paris, 2015).
12. Clegg, J. P., Guest, J. F., Lehman, A. & Smith, A. F., The annual cost of dry eye syndrome in France, Germany, Italy, Spain, Sweden and the United Kingdom among patients managed by ophthalmologists, *Ophthalmic epidemiology* **13**, 263–274 (2006).

-
13. Sergheraert, L., Le syndrome de l'œil sec, une pathologie en forte progression, *Actualités Pharmaceutiques* **61**, 35–38 (2022).
 14. Wolffsohn, J. S. *et al.*, Demographic and lifestyle risk factors of dry eye disease subtypes: a cross-sectional study, *The Ocular Surface* **21**, 58–63 (2021).
 15. Tear film mucins: front line defenders of the ocular surface; comparison with airway and gastrointestinal tract mucins.
 16. Stapleton, F. *et al.*, Tfos dewes ii epidemiology report, *The ocular surface* **15**, 334–365 (2017).
 17. Schaumberg, D. A. *et al.*, The international workshop on meibomian gland dysfunction: report of the subcommittee on the epidemiology of, and associated risk factors for, MGD, *Investigative ophthalmology & visual science* **52**, 1994–2005 (2011).
 18. Shiboski, S. *et al.*, American College of Rheumatology classification criteria for Sjögren's syndrome: a data-driven, expert consensus approach in the Sjögren's International Collaborative Clinical Alliance cohort, *Arthritis care & research* **64**, 475–487 (2012).
 19. *Necessity* Dec. 2021, <https://www.necessity-h2020.eu/>.
 20. Van Bijsterveld, O., Diagnostic tests in the sicca syndrome, *Archives of ophthalmology* **82**, 10–14 (1969).
 21. Senchyna, M. & Wax, M. B., Quantitative assessment of tear production: A review of methods and utility in dry eye drug discovery, *Journal of Ocular Biology, Diseases, and Informatics* **1**, 1–6, ISSN: 1936-8445, (2019-12-23) (2008).
 22. Savini, G., The challenge of dry eye diagnosis, *Clinical Ophthalmology*, 31, ISSN: 1177-5483, (2019-12-23) (2008).
 23. Simpson, T. L., Situ, P., Jones, L. W. & Fonn, D., Dry Eye Symptoms Assessed by Four Questionnaires: *Optometry and Vision Science* **85**, E692–E699, ISSN: 1040-5488, (2019-12-23) (2008).
 24. Sweeney, D. F., Millar, T. J. & Raju, S. R., Tear film stability: A review, *Experimental Eye Research* **117**, 28–38, ISSN: 00144835, (2019-12-23) (2013).
 25. Latkany, R., Miller, D. & Zeev, M. S.-B., Diagnosis of dry eye disease and emerging technologies, *Clinical Ophthalmology*, 581, ISSN: 1177-5483, (2019-12-23) (2014-03).

-
26. Garaszczuk, I. K., Montes Mico, R., Iskander, D. R. & Expósito, A. C., The tear turnover and tear clearance tests – a review, *Expert Review of Medical Devices* **15**, 219–229, ISSN: 1743-4440, 1745-2422, (2019-12-23) (2018-03).
 27. Begley, C. *et al.*, Review and analysis of grading scales for ocular surface staining, *The ocular surface* (2019).
 28. Schirmer, O., Studies on the physiology and pathology of the secretion and drainage of tears. *Albrecht von Graefes Arch Klin Ophthalmol* **56**, 197–291 (1903).
 29. Clinch, T. E., Benedetto, D. A., Felberg, N. T. & Laibson, P. R., Schirmer’s test: a closer look, *Archives of ophthalmology* **101**, 1383–1386 (1983).
 30. Cho Pauline, Y. M., Schirmer test. I. A review. *Optometry and vision science: official publication of the American Academy of Optometry* **70**, 152–156 (1993).
 31. Henderson, J. W. & Prough, W. A., Influence of age and sex on flow of tears, *Archives of Ophthalmology* **43**, 224–231 (1950).
 32. Prause, J. U., Frost-Larsen, K., Isager, H. & Manthorpe, R., TEAR ABSORPTION INTO THE FILTER-PAPER STRIP USED IN THE SCHIRMER-I-TEST: A methodological study and a critical survey, *Acta ophthalmologica* **60**, 70–78 (1982).
 33. Pandher, K., Mengher, L., Duerden, J. & Bron, A., Effect of meibomian oils on Schirmer tear test, *Acta ophthalmologica* **63**, 695–697 (1985).
 34. Wright, J. & Meger, G., A review of the Schirmer test for tear production, *Archives of Ophthalmology* **67**, 564–565 (1962).
 35. Hanson, J., Fikentscher, R. & Roseburg, B., Schirmer test of lacrimation: Its clinical importance, *Archives of Otolaryngology* **101**, 293–295 (1975).
 36. Patel, S., Farrell, J. & Bevan, R., Reliability and variability of the Schirmer test, *Optician* **194**, 12–14 (1987).
 37. Halberg, G. P. & Berens, C., Standardized Schirmer tear test kit, *American journal of ophthalmology* **51**, 840–842 (1961).
 38. Jones, L. T., The lacrimal secretory system and its treatment, *American journal of ophthalmology* **62**, 47–60 (1966).
 39. Doughman, D. J., Clinical tests, *International ophthalmology clinics* **13**, 199–220 (1973).

-
40. Shapiro, A. & Merin, S., Schirmer test and break-up time of tear film in normal subjects, *American journal of ophthalmology* **88**, 752–757 (1979).
 41. Nichols, K. K., Nichols, J. J. & Mitchell, G. L., The lack of association between signs and symptoms in patients with dry eye disease, *Cornea* **23**, 762–770 (2004).
 42. Nelson, P., A shorter Schirmer tear test, *Optom Mon* **73**, 568–9 (1982).
 43. Bawazeer, A. M. & Hodge, W. G., One-minute Schirmer test with anesthesia, *Cornea* **22**, 285–287 (2003).
 44. Kurihashi, K., Yanagihara, N. & Honda, Y., A modified Schirmer test: the fine-thread method for measuring lacrimation, *Journal of Pediatric Ophthalmology and Strabismus* **14**, 390–397 (1977).
 45. Hamano, H., A new method for measuring tears. *CLAO J* **9**, 281–289 (1983).
 46. Prabhasawat, P. & Tseng, S. C., Frequent association of delayed tear clearance in ocular irritation, *British journal of ophthalmology* **82**, 666–675 (1998).
 47. Pflugfelder, S. C. *et al.*, Evaluation of subjective assessments and objective diagnostic tests for diagnosing tear-film disorders known to cause ocular irritation. *Cornea* **17**, 38–56 (1998).
 48. Xu, K.-P., Yagi, Y., Toda, I. & Tsubota, K., Tear function index: a new measure of dry eye, *Archives of ophthalmology* **113**, 84–88 (1995).
 49. Guillon, J.-P., Non-invasive tearscope plus routine for contact lens fitting, *Contact Lens and Anterior Eye* **21**, S31–S40 (1998).
 50. Dogru, M. *et al.*, Strip Meniscometry: A New and Simple Method of Tear Meniscus Evaluation, *Investigative Ophthalmology & Visual Science* **47**, 1895, ISSN: 1552-5783, (2019-12-23) (2006-05).
 51. Lee, W. B. & Mannis, M. J., Historical Concepts of Ocular Surface Disease, 3–10 (2013).
 52. Argüeso, P., Tisdale, A., Spurr-Michaud, S., Sumiyoshi, M. & Gipson, I. K., Mucin characteristics of human corneal-limbal epithelial cells that exclude the rose bengal anionic dye, *Investigative ophthalmology & visual science* **47**, 113–119 (2006).
 53. Pflüger, Zu Ernährung der cornea, *Klin. Monatsbl. Augenheilk* **20**, 69–81 (1882).

-
54. Bron, A., Argüeso, P., Irkeç, M. & Bright, F., Clinical staining of the ocular surface: Mechanisms and interpretations, *Progress in Retinal and Eye Research* **44**, 36–61, ISSN: 13509462, (2019-12-23) (2015-01).
 55. Woods, J., Varikooty, J., Fonn, D. & Jones, L. W., A novel scale for describing corneal staining, *Clinical Ophthalmology (Auckland, NZ)* **12**, 2369 (2018).
 56. Norn, M., DESICCATION OF THE PRECORNEAL FILM: I. Corneal Wetting-Time, *Acta ophthalmologica* **47**, 865–880 (1969).
 57. Morgan, P. B., Tullo, A. B. & Efron, N., Infrared thermography of the tear film in dry eye, *Eye* **9**, 615–618, ISSN: 0950-222X, 1476-5454 (1995-09).
 58. Fujishima, H., Toda, I., Yamada, M., Sato, N. & Tsubota, K., Corneal temperature in patients with dry eye evaluated by infrared radiation thermometry. *British Journal of Ophthalmology* **80**, 29–32, ISSN: 0007-1161, (2019-12-23) (1996-01).
 59. Hirji, N., Patel, S. & Callander, M., Human tear film pre-rupture phase time (TP-RPT)-A non-invasive technique for evaluating the pre-corneal tear film using a novel keratometer mire, *Ophthalmic and Physiological Optics* **9**, 139–142 (1989).
 60. Nichols, K. K. *et al.*, The international workshop on meibomian gland dysfunction: executive summary, *Investigative Ophthalmology Visual Science* **52**, 1922–1929, ISSN: 1552-5783 (Mar. 2011).
 61. Xiao, J. *et al.*, Diagnostic Test Efficacy of Meibomian Gland Morphology and Function, *Scientific Reports* **9**, 17345 (2019).
 62. Lemp, M. A. & Foulks, G. N., The definition and classification of dry eye disease, *Ocul Surf* **5**, 75–92 (2007).
 63. Shiraishi, A. & Sakane, Y., Assessment of Dry Eye Symptoms: Current Trends and Issues of Dry Eye Questionnaires in Japan, *Investigative ophthalmology & visual science* **59**, DES23–DES28 (2018).
 64. Smith, J. A., The epidemiology of dry eye disease, *Acta Ophthalmologica Scandinavica* **85** (2007).
 65. Begley, C. G., Caffery, B., Chalmers, R. L., Mitchell, G. L., *et al.*, Use of the dry eye questionnaire to measure symptoms of ocular irritation in patients with aqueous tear deficient dry eye, *Cornea* **21**, 664–670 (2002).

-
66. Vitale, S., Goodman, L. A., Reed, G. F. & Smith, J. A., Comparison of the NEI-VFQ and OSDI questionnaires in patients with Sjögren's syndrome-related dry eye, *Health and quality of life outcomes* **2**, 44 (2004).
 67. Chalmers, R. L., Begley, C. G. & Caffery, B., Validation of the 5-Item Dry Eye Questionnaire (DEQ-5): Discrimination across self-assessed severity and aqueous tear deficient dry eye diagnoses, *Contact Lens and Anterior Eye* **33**, 55–60 (2010).
 68. Grubbs Jr, J. R., Tolleson-Rinehart, S., Huynh, K. & Davis, R. M., A review of quality of life measures in dry eye questionnaires, *Cornea* **33**, 215 (2014).
 69. Tomlinson, A., Khanal, S., Ramaesh, K., Diaper, C. & McFadyen, A., Tear film osmolarity: determination of a referent for dry eye diagnosis, *Investigative ophthalmology & visual science* **47**, 4309–4315 (2006).
 70. Suzuki, M. *et al.*, Tear Osmolarity as a Biomarker for Dry Eye Disease Severity, *Investigative Ophthalmology & Visual Science* **51**, 4557, ISSN: 1552-5783, (2010-12-23) (2010-09).
 71. Karns, K. & Herr, A. E., OPHTHALMOLOGIST-ON-A-CHIP: FULLY INTEGRATED MICROFLUIDIC TEAR OSMOLARITY AND PROTEIN BIOMARKER QUANTIFICATION FOR DRY EYE STRATIFICATION, *International Conference on Miniaturized Systems for Chemistry and Life Sciences* (2011).
 72. Pucker, A. D. & Nichols, J. J., Analysis of Meibum and Tear Lipids, *The Ocular Surface* **10**, 230–250, ISSN: 15420124, (2010-12-23) (2012).
 73. Srinivasan, S., Thangavelu, M., Zhang, L., Green, K. B. & Nichols, K. K., iTRAQ Quantitative Proteomics in the Analysis of Tears in Dry Eye Patients, *Investigative Ophthalmology & Visual Science* **53**, 5052, ISSN: 1552-5783 (2012).
 74. Brewitt, H. & Sistani, F., Dry eye disease: the scale of the problem, *Survey of ophthalmology* **45**, S199–S202 (2001).
 75. Zeev, M. S.-B., Miller, D. D. & Latkany, R., Diagnosis of dry eye disease and emerging technologies, *Clinical Ophthalmology (Auckland, NZ)* **8**, 581 (2014).
 76. Brooke Henson, E., Tracy Schroeder Swartz, O., Marguerite McDonald, M., Eric Donnenfeld, M. & Crystal Brimer, O., *Diagnosis and overlooking asymptomatic patients with DED* June 2021, <https://www.optometrytimes.com/view/diagnosis-and-overlooking-asymptomatic-patients-with-ded>.

-
77. Alzubaidi, L. *et al.*, Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions, *Journal of big Data* **8**, 1–74 (2021).
 78. Young, T., Hazarika, D., Poria, S. & Cambria, E., Recent trends in deep learning based natural language processing, *ieeE Computational intelligence magazine* **13**, 55–75 (2018).
 79. Adeel, A., Gogate, M. & Hussain, A., Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments, *Information Fusion* **59**, 163–170 (2020).
 80. Tian, H., Chen, S.-C. & Shyu, M.-L., Evolutionary programming based deep learning feature selection and network construction for visual data classification, *Information systems frontiers* **22**, 1053–1066 (2020).
 81. says: R. & says: S. B., *Ai vs machine learning vs deep learning* Jan. 2022, <https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/>.
 82. Research, V. M., *2022 statistics: U.S. and Global Artificial Intelligence (AI) in healthcare market size will surpass USD 95.65 billion at 46.1% CAGR growth: Vantage Market Research* Jan. 2022, <https://www.globenewswire.com/news-release/2022/01/17/2367615/0/en/2022-Statistics-U-S-and-Global-Artificial-Intelligence-AI-in-Healthcare-Market-Size-Will-Surpass-USD-95-65-Billion-at-46-1-CAGR-Growth-Vantage-Market-Research.html>.
 83. Serte, S., Serener, A. & Al-Turjman, F., Deep learning in medical imaging: A brief review, *Transactions on Emerging Telecommunications Technologies*, e4080 (2020).
 84. Kim, M. *et al.*, Deep learning in medical imaging, *Neurospine* **16**, 657 (2019).
 85. Anaya-Isaza, A., Mera-Jiménez, L. & Zequera-Diaz, M., An overview of deep learning in medical imaging, *Informatics in Medicine Unlocked* **26**, 100723 (2021).
 86. Lee, J.-G. *et al.*, Deep learning in medical imaging: general overview, *Korean journal of radiology* **18**, 570–584 (2017).
 87. Bishop, C. M. & Nasrabadi, N. M., *Pattern recognition and machine learning* **4** (Springer, 2006).
 88. Castiglioni, I. *et al.*, AI applications to medical images: From machine learning to deep learning, *Physica Medica* **83**, 9–24 (2021).

-
89. Ning, X. *et al.*, A review of research on co-training, *Concurrency and computation: practice and experience*, e6276 (2021).
 90. Cheplygina, V., de Bruijne, M. & Pluim, J. P., Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, *Medical image analysis* **54**, 280–296 (2019).
 91. Wang, G., Wang, G. & Wu, Q. M. J., *Guide to three dimensional structure and motion factorization* ISBN: 978-0-85729-046-5 (Springer, 2011).
 92. Laveau, S. & Faugeras, O., *Oriented projective geometry for computer vision in European Conference on Computer Vision* (1996), 147–156.
 93. Wang, G. & Wu, Q. J., *Guide to three dimensional structure and motion factorization* (Springer, 2011).
 94. Rother, C., *Multi-view reconstruction and camera recovery using a real or virtual reference plane* PhD thesis (Numerisk analys och datalogi, 2003).
 95. Saha, S., *A comprehensive guide to Convolutional Neural Networks-the eli5 way* Dec. 2018, <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>.
 96. Santosh, K., AI-driven tools for coronavirus outbreak: need of active learning and cross-population train/test models on multitudinal/multimodal data, *Journal of medical systems* **44**, 1–5 (2020).
 97. Santosh, K. & Ghosh, S., Covid-19 imaging tools: How big data is big?, *Journal of Medical Systems* **45**, 1–8 (2021).
 98. Das, D., Santosh, K. & Pal, U., Truncated inception net: COVID-19 outbreak screening using chest X-rays, *Physical and engineering sciences in medicine* **43**, 915–925 (2020).
 99. Santosh, K., COVID-19 prediction models and unexploited data, *Journal of medical systems* **44**, 1–4 (2020).
 100. Mukherjee, H. *et al.*, Deep neural network to detect COVID-19: one architecture for both CT Scans and Chest X-rays, *Applied Intelligence* **51**, 2777–2789 (2021).
 101. Su, T.-Y., Ting, P.-J., Chang, S.-W. & Chen, D.-Y., Superficial Punctate Keratitis Grading for Dry Eye Screening using Deep Convolutional Neural Networks, *IEEE Sensors Journal*, 1–1, (2019-12-23) (2019).

-
102. Chun, Y. S., Yoon, W. B., Kim, K. G. & Park, I. K., Objective Assessment of Corneal Staining Using Digital Image Analysis, *Investigative Ophthalmology & Visual Science* **55**, 7896–7903, ISSN: 0146-0404, (2019-12-23) (2014-12).
 103. Pellegrini, M. *et al.*, Assessment of Corneal Fluorescein Staining in Different Dry Eye Subtypes Using Digital Image Analysis, *Translational Vision Science & Technology* **8**, 34–34 (2019).
 104. Rodriguez, J. D. *et al.*, Automated Grading System for Evaluation of Superficial Punctate Keratitis Associated With Dry Eye, *Investigative Ophthalmology & Visual Science* **56**, (2019-12-23) (2015-04).
 105. Su, T.-Y., Liu, Z.-Y. & Chen, D.-Y., Tear Film Break-Up Time Measurement Using Deep Convolutional Neural Networks for Screening Dry Eye Disease, *IEEE Sensors Journal* **18**, 6857–6862 (2018).
 106. Wang, J., Yeh, T. N., Chakraborty, R., Stella, X. Y. & Lin, M. C., A Deep Learning Approach for Meibomian Gland Atrophy Evaluation in Meibography Images, *Translational Vision Science & Technology* **8**, 37–37 (2019).
 107. Prabhu, S. M., Chakiat, A., Shashank, S., Vunnava, K. P. & Shetty, R., Deep learning segmentation and quantification of Meibomian glands, *Biomedical Signal Processing and Control* **57**, 101776 (2020).
 108. Fraundorfer, F. & Scaramuzza, D., Visual odometry: Part i: The first 30 years and fundamentals, *IEEE Robotics and Automation Magazine* **18**, 80–92 (2011).
 109. Cetinkaya, G., *Visual odometry vs. visual slam vs. structure-from-motion* June 2022, <https://guvencetinkaya.medium.com/visual-odometry-vs-visual-slam-cdda75df592>.
 110. Aqel, M. O., Marhaban, M. H., Saripan, M. I. & Ismail, N. B., Review of visual odometry: types, approaches, challenges, and applications, *SpringerPlus* **5**, 1–26 (2016).
 111. Nistér, D., Naroditsky, O. & Bergen, J., *Visual odometry in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* **1** (2004), I–I.
 112. Howard, A., *Real-time stereo visual odometry for autonomous ground vehicles in 2008 IEEE/RSJ international conference on intelligent robots and systems* (2008), 3946–3952.

-
113. Lowe, D. G., Distinctive image features from scale-invariant keypoints, *International journal of computer vision* **60**, 91–110 (2004).
 114. Bay, H., Tuytelaars, T. & Gool, L. V., *Surf: Speeded up robust features in European conference on computer vision* (2006), 404–417.
 115. Rublee, E., Rabaud, V., Konolige, K. & Bradski, G., *ORB: An efficient alternative to SIFT or SURF in 2011 International conference on computer vision* (2011), 2564–2571.
 116. Mur-Artal, R. & Tardós, J. D., Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras, *IEEE transactions on robotics* **33**, 1255–1262 (2017).
 117. Rosten, E. & Drummond, T., *Machine learning for high-speed corner detection in European conference on computer vision* (2006), 430–443.
 118. Calonder, M., Lepetit, V., Strecha, C. & Fua, P., *Brief: Binary robust independent elementary features in European conference on computer vision* (2010), 778–792.
 119. Ma, T. *et al.*, *Robust image matching via ORB feature and VFC for mismatch removal in MIPPR 2017: Pattern Recognition and Computer Vision* **10609** (2018), 69–74.
 120. Lovegrove, S., Davison, A. J. & Ibanez-Guzmán, J., *Accurate visual odometry from a rear parking camera in 2011 IEEE Intelligent Vehicles Symposium (IV)* (2011), 788–793.
 121. Newcombe, R. A., Lovegrove, S. J. & Davison, A. J., *DTAM: Dense tracking and mapping in real-time in 2011 international conference on computer vision* (2011), 2320–2327.
 122. Engel, J., Koltun, V. & Cremers, D., Direct sparse odometry, *IEEE transactions on pattern analysis and machine intelligence* **40**, 611–625 (2017).
 123. Stühmer, J., Gumhold, S. & Cremers, D., *Real-time dense geometry from a handheld camera in Joint Pattern Recognition Symposium* (2010), 11–20.
 124. Pizzoli, M., Forster, C. & Scaramuzza, D., *REMODE: Probabilistic, monocular dense reconstruction in real time in 2014 IEEE international conference on robotics and automation (ICRA)* (2014), 2609–2616.
 125. Engel, J., Schöps, T. & Cremers, D., *LSD-SLAM: Large-scale direct monocular SLAM in European conference on computer vision* (2014), 834–849.

-
126. Lu, F., Zhou, B., Zhang, Y. & Zhao, Q., Real-time 3D scene reconstruction with dynamically moving object using a single depth camera, *The Visual Computer* **34**, 753–763 (2018).
 127. Jin, H., Favaro, P. & Soatto, S., A semi-direct approach to structure from motion, *The Visual Computer* **19**, 377–394 (2003).
 128. Zhou, Y., Yan, F. & Zhou, Z., Handling pure camera rotation in semi-dense monocular SLAM, *The Visual Computer* **35**, 123–132 (2019).
 129. Silveira, G., Malis, E. & Rives, P., An efficient direct approach to visual SLAM, *IEEE transactions on robotics* **24**, 969–979 (2008).
 130. Forster, C., Pizzoli, M. & Scaramuzza, D., *SVO: Fast semi-direct monocular visual odometry in 2014 IEEE international conference on robotics and automation (ICRA)* (2014), 15–22.
 131. Baker, S. & Matthews, I., Lucas-kanade 20 years on: A unifying framework, *International journal of computer vision* **56**, 221–255 (2004).
 132. Lu, R., Zhu, F., Wu, Q. & Fu, X., Search inliers based on redundant geometric constraints, *The Visual Computer* **36**, 253–266 (2020).
 133. Macario Barros, A., Michel, M., Moline, Y., Corre, G. & Carrel, F., A comprehensive survey of visual slam algorithms, *Robotics* **11**, 24 (2022).
 134. Azzam, R., Taha, T., Huang, S. & Zweiri, Y., Feature-based visual simultaneous localization and mapping: A survey, *SN Applied Sciences* **2**, 1–24 (2020).
 135. Davison, A. J., Reid, I. D., Molton, N. D. & Stasse, O., MonoSLAM: Real-time single camera SLAM, *IEEE transactions on pattern analysis and machine intelligence* **29**, 1052–1067 (2007).
 136. Klein, G. & Murray, D., *Parallel tracking and mapping for small AR workspaces in 2007 6th IEEE and ACM international symposium on mixed and augmented reality* (2007), 225–234.
 137. Tateno, K., Tombari, F., Laina, I. & Navab, N., *Cnn-slam: Real-time dense monocular slam with learned depth prediction in Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 6243–6252.

-
138. Nyimbili, P. H., Demirel, H., Seker, D. & Erden, T., *Structure from motion (sfm)-approaches and applications in Proceedings of the international scientific conference on applied sciences, Antalya, Turkey* (2016), 27–30.
 139. Mohr, R., Quan, L. & Veillon, F., Relative 3D reconstruction using multiple uncalibrated images, *The International Journal of Robotics Research* **14**, 619–632 (1995).
 140. Beardsley, P., Torr, P. & Zisserman, A., *3D model acquisition from extended image sequences in European conference on computer vision* (1996), 683–695.
 141. Fitzgibbon, A. W. & Zisserman, A., *Automatic camera recovery for closed or open image sequences in European conference on computer vision* (1998), 311–326.
 142. Pollefeys, M. *et al.*, Visual modeling with a hand-held camera, *International Journal of Computer Vision* **59**, 207–232 (2004).
 143. Dellaert, F., Seitz, S. M., Thorpe, C. E. & Thrun, S., *Structure from motion without correspondence in Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)* **2** (2000), 557–564.
 144. Agarwal, S. *et al.*, Building rome in a day, *Communications of the ACM* **54**, 105–112 (2011).
 145. Crandall, D., Owens, A., Snavely, N. & Huttenlocher, D., *Discrete-continuous optimization for large-scale structure from motion in CVPR 2011* (2011), 3001–3008.
 146. Wilson, K. & Snavely, N., *Robust global translations with 1dsfm in European conference on computer vision* (2014), 61–75.
 147. Sweeney, C., Sattler, T., Hollerer, T., Turk, M. & Pollefeys, M., *Optimizing the viewing graph for structure-from-motion in Proceedings of the IEEE international conference on computer vision* (2015), 801–809.
 148. Gherardi, R., Farenzena, M. & Fusiello, A., *Improving the efficiency of hierarchical structure-and-motion in 2010 IEEE computer society conference on computer vision and pattern recognition* (2010), 1594–1600.
 149. Snavely, N., Seitz, S. M. & Szeliski, R., *in ACM siggraph 2006 papers* 835–846 (2006).
 150. Wu, C., *Towards linear-time incremental structure from motion in 2013 International Conference on 3D Vision-3DV 2013* (2013), 127–134.

-
151. Frahm, J.-M. *et al.*, *Building rome on a cloudless day in European conference on computer vision* (2010), 368–381.
 152. Shalaby, A., Elmogy, M. & El-Fetouh, A. A., Algorithms and applications of structure from motion (SFM): A survey, *Algorithms* **6** (2017).
 153. Fuentes-Pacheco, J., Ruiz-Ascencio, J. & Rendón-Mancha, J. M., Visual simultaneous localization and mapping: a survey, *Artificial intelligence review* **43**, 55–81 (2015).
 154. Cadena, C. *et al.*, Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age, *IEEE Transactions on robotics* **32**, 1309–1332 (2016).
 155. Huang, S. & Dissanayake, G., A critique of current developments in simultaneous localization and mapping, *International Journal of Advanced Robotic Systems* **13**, 1729881416669482 (2016).
 156. Saeedi, S., Trentini, M., Seto, M. & Li, H., Multiple-robot simultaneous localization and mapping: A review, *Journal of Field Robotics* **33**, 3–46 (2016).
 157. Li, R., Wang, S. & Gu, D., Ongoing evolution of visual slam from geometry to deep learning: Challenges and opportunities, *Cognitive Computation* **10**, 875–889 (2018).
 158. Saputra, M. R. U., Markham, A. & Trigoni, N., Visual SLAM and structure from motion in dynamic environments: A survey, *ACM Computing Surveys (CSUR)* **51**, 1–36 (2018).
 159. Sualeh, M. & Kim, G.-W., Simultaneous localization and mapping in the epoch of semantics: a survey, *International Journal of Control, Automation and Systems* **17**, 729–742 (2019).
 160. Wang, K., Ma, S., Chen, J., Ren, F. & Lu, J., Approaches challenges and applications for deep visual odometry toward to complicated and emerging areas, *IEEE Transactions on Cognitive and Developmental Systems* (2020).
 161. Zhao, C., Sun, Q., Zhang, C., Tang, Y. & Qian, F., Monocular depth estimation based on deep learning: An overview, *Science China Technological Sciences* **63**, 1612–1627 (2020).
 162. Geiger, A., Lenz, P., Stiller, C. & Urtasun, R., Vision meets robotics: The kitti dataset, *The International Journal of Robotics Research* **32**, 1231–1237 (2013).

-
163. Geiger, A., Lenz, P. & Urtasun, R., *Are we ready for autonomous driving? the kitti vision benchmark suite in 2012 IEEE conference on computer vision and pattern recognition* (2012), 3354–3361.
 164. Mayer, N. *et al.*, *A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation in Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 4040–4048.
 165. Zhao, C., Tang, Y., Sun, Q. & Vasilakos, A. V., *Deep direct visual odometry, IEEE transactions on intelligent transportation systems* (2021).
 166. Eigen, D., Puhrsch, C. & Fergus, R., *Depth map prediction from a single image using a multi-scale deep network, Advances in neural information processing systems* **27** (2014).
 167. Chen, X., Ma, H., Wan, J., Li, B. & Xia, T., *Multi-view 3d object detection network for autonomous driving in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2017), 1907–1915.
 168. Wang, P. *et al.*, *Understanding convolution for semantic segmentation in 2018 IEEE winter conference on applications of computer vision (WACV)* (2018), 1451–1460.
 169. Chang, M.-F. *et al.*, *Argoverse: 3d tracking and forecasting with rich maps in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 8748–8757.
 170. Xue, F. *et al.*, *Beyond tracking: Selecting memory and refining poses for deep visual odometry in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), 8575–8583.
 171. Clark, R., Wang, S., Wen, H., Markham, A. & Trigoni, N., *Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem in Proceedings of the AAAI Conference on Artificial Intelligence* **31** (2017).
 172. Cordts, M. *et al.*, *The cityscapes dataset for semantic urban scene understanding in Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 3213–3223.
 173. Godard, C., Mac Aodha, O. & Brostow, G. J., *Unsupervised monocular depth estimation with left-right consistency in Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 270–279.

-
174. Yin, Z. & Shi, J., *Geonet: Unsupervised learning of dense depth, optical flow and camera pose in Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 1983–1992.
 175. Bian, J. *et al.*, Unsupervised scale-consistent depth and ego-motion learning from monocular video, *Advances in neural information processing systems* **32** (2019).
 176. Zhou, T., Brown, M., Snavely, N. & Lowe, D. G., *Unsupervised learning of depth and ego-motion from video in Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 1851–1858.
 177. Brahim, S., Ben Aoun, N., Ben Amar, C., Benoit, A. & Lambert, P., Multiscale fully convolutional densenet for semantic segmentation (2018).
 178. Szeliski, R., *Prediction error as a quality metric for motion and stereo in Proceedings of the Seventh IEEE International Conference on Computer Vision* **2** (1999), 781–788.
 179. Tzutalin, *LabelImg* <https://github.com/tzutalin/labelImg>, 2015.
 180. Guerre, A., *Champ visuel augmenté pour l'exploration vidéo de la rétine* PhD thesis (Université de Bretagne occidentale-Brest, 2019).
 181. Guerre, A., Lamard, M., Conze, P.-H., Cochener, B. & Quellec, G., *Optical flow estimation in ocular endoscopy videos using flownet on simulated endoscopy data in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (2018), 1463–1466.
 182. Rau, A. *et al.*, Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy, *International journal of computer assisted radiology and surgery* **14**, 1167–1176 (2019).
 183. Gordon, A., Li, H., Jonschkowski, R. & Angelova, A., *Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras in Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), 8977–8986.
 184. Alhashim, I. & Wonka, P., High quality monocular depth estimation via transfer learning, *arXiv preprint arXiv:1812.11941* (2018).
 185. Zheng, C., Cham, T.-J. & Cai, J., *T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks in Proceedings of the European conference on computer vision (ECCV)* (2018), 767–783.

-
186. Benalcazar, D. P., Zambrano, J. E., Bastias, D., Perez, C. A. & Bowyer, K. W., A 3D iris scanner from a single image using convolutional neural networks, *IEEE Access* **8**, 98584–98599 (2020).
187. Ozyoruk, K. B. *et al.*, EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos, *Medical image analysis* **71**, 102058 (2021).
188. Casser, V., Pirk, S., Mahjourian, R. & Angelova, A., *Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos in Proceedings of the AAAI conference on artificial intelligence* **33** (2019), 8001–8008.
189. Clément, G., Mac, A. O., Michael, F. & Gabriel, B., Digging Into Self-Supervised Monocular Depth Estimation, *arXiv preprint arXiv:1806.01260* (2018).
190. Bréhéret, A., *Pixel Annotation Tool* <https://github.com/abreheret/PixelAnnotationTool>, 2017.
191. Iakubovskii, P., *Segmentation Models Pytorch* https://github.com/qubvel/segmentation_models.pytorch, 2019.
192. Ronneberger, O., Fischer, P. & Brox, T., *U-net: Convolutional networks for biomedical image segmentation in International Conference on Medical image computing and computer-assisted intervention* (2015), 234–241.
193. Lin, T.-Y. *et al.*, *Feature pyramid networks for object detection in Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), 2117–2125.
194. Bloice, M. D., Roth, P. M. & Holzinger, A., Biomedical image augmentation using Augmentor, *Bioinformatics* **35**, 4522–4524 (2019).
195. Dosovitskiy, A. *et al.*, *Flownet: Learning optical flow with convolutional networks in Proceedings of the IEEE international conference on computer vision* (2015), 2758–2766.
196. Godard, C., Mac Aodha, O., Firman, M. & Brostow, G. J., Digging into Self-Supervised Monocular Depth Prediction (Oct. 2019).
197. Jaderberg, M., Simonyan, K., Zisserman, A., *et al.*, Spatial transformer networks, *Advances in neural information processing systems* **28** (2015).

-
198. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P., Image quality assessment: from error visibility to structural similarity, *IEEE transactions on image processing* **13**, 600–612 (2004).
 199. Mahjourian, R., Wicke, M. & Angelova, A., *Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints in Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 5667–5675.
 200. Besl, P. J. & McKay, N. D., *Method for registration of 3-D shapes in Sensor fusion IV: control paradigms and data structures* **1611** (1992), 586–606.
 201. Bekerman, I., Gottlieb, P. & Vaiman, M., Variations in eyeball diameters of the healthy adults, *Journal of ophthalmology* **2014** (2014).
 202. *Anatomy of the eye* <https://diagnosis101.welchallyn.com/ophthalmoscopy/educational-topics/anatomy-of-the-eye/>.
 203. Davies, E. R., *Machine vision: theory, algorithms, practicalities* (Elsevier, 2004).
 204. Jekel, C. F., *Digital Image Correlation on Steel Ball* 83–87, <https://hdl.handle.net/10019.1/98627> (2016).
 205. Welch, G., Bishop, G., *et al.*, An introduction to the Kalman filter (1995).
 206. Bron, A., Reflections on the tears, *Eye* **11**, 583 (1997).
 207. Smith, L. N. & Topin, N., *Super-convergence: Very fast training of neural networks using large learning rates in Artificial intelligence and machine learning for multi-domain operations applications* **11006** (2019), 369–386.
 208. Loshchilov, I. & Hutter, F., Sgdr: Stochastic gradient descent with warm restarts, *arXiv preprint arXiv:1608.03983* (2016).
 209. Farrand, K. F., Fridman, M., Stillman, I. Ö. & Schaumberg, D. A., Prevalence of diagnosed dry eye disease in the United States among adults aged 18 years and older, *American journal of ophthalmology* **182**, 90–98 (2017).
 210. Wadell, H., Volume, shape, and roundness of rock particles, *The Journal of Geology* **40**, 443–451 (1932).
 211. Cruz-Matias, I. *et al.*, Sphericity and roundness computation for particles using the extreme vertices model, *Journal of computational science* **30**, 28–40 (2019).

Titre : Quantification automatique de la sécheresse oculaire par intelligence artificielle au cours du syndrome de Sjögren

Mot clés : Sécheresse oculaire, Classification automatique, Apprentissage profond, Conformité de structure

Résumé : Le syndrome de Sjögren est une maladie du système immunitaire dont les deux symptômes communs sont la sécheresse des yeux et de la bouche. La gêne occasionnée par les symptômes de sécheresse oculaire affecte la vie quotidienne, entraîne une diminution de 30% des activités et touche 95% des patients atteints du syndrome de Sjögren [1]. La sécheresse oculaire est également un trouble multifactoriel indépendant dont la prévalence peut atteindre 50% [2]. L'inflammation de la surface oculaire entraîne une gêne, une fatigue et, globalement, une baisse de la qualité de vie [2, 3]. Les thérapies traditionnelles permettent de gérer les symptômes et d'éviter les dommages permanents. Il est donc essentiel de classer et de suivre l'évolution de la DED. Les méthodes existantes qui permettent de diagnostiquer et de quantifier les DED présentent des inconvénients communs : reproductibilité, invasivité et imprécision. Nous avons passé en revue les méthodes classiques et celles qui intègrent l'automatisation pour mesurer l'étendue de la DED : [4]. Cette étude a montré que la DED n'a pas encore bénéficié de ce que l'intelligence artificielle (IA) a à offrir. En utilisant des examens de la surface oculaire à la lampe à fente, nous avons cherché à améliorer la quantification du score d'Oxford [5]. La méthode que nous proposons utilise l'apprentissage non supervisé pour recalibrer les images des examens dans un système de coordonnées commun. En apprenant simultanément le mouvement de la caméra et la pro-

fondeur, nous sommes en mesure de suivre la surface oculaire en 3D, de compenser le mouvement de l'œil et de visualiser l'œil entier. La source lumineuse fixée à la caméra constitue un défi et une perturbation lors de l'apprentissage de l'égomotion. Ce problème a été résolu par la segmentation sémantique et l'ajout d'un nouveau signal de supervision : la perte de reconstruction sémantique. Nous avons également utilisé la forme de l'œil comme une connaissance préalable que nous pouvons inclure comme une contrainte. Ceci a été mis en œuvre par une perte d'ajustement de forme ; les formes étant deux sphères se croisant l'une l'autre. Notre recalage a montré une amélioration quantitative et qualitative avec chaque contribution. Nous avons également calculé la fiabilité inter-juges des annotations des points ponctuels (zones endommagées). Notre méthode s'est rapprochée le plus de ce qui peut être considéré comme une erreur humaine. La méthode de recalage proposée a également été utilisée pour une tâche de prétraitement, la sélection des images. Une fois appliquée à la classification automatique du score d'Oxford, notre méthode a également amélioré les résultats. Cette amélioration valide le fait que les fortes variations de couleur/illumination présentes dans les examens constituent une perturbation pour toute tâche d'apprentissage profond. Nous avons surmonté ce problème dans les deux tâches grâce à nos contributions et à la méthode proposée.

Title: Automatic quantification of ocular dryness by artificial intelligence in the context of Sjögren's syndrome

Keywords: Dry eye disease, Automated grading, Deep learning, Shape fitting

Abstract: Sjögren's syndrome is an immune system disorder with two common symptoms, dry eyes and a dry mouth. The discomfort of dry eye symptoms affects daily lives, results in 30% activity impairment and affects 95% of Sjögren patients [1]. Dry eye disease (DED) is also an independent multifactorial disorder with a prevalence of up to 50% [2]. The ocular surface inflammation causes discomfort, fatigue and overall, a lower quality of life [2, 3]. Traditional therapies help manage the symptoms and avoid permanent damage. Hence, it is pivotal to grade and follow the development of DED. A common drawback in existing methods that diagnose and quantify DED is reproducibility, invasivity and inaccuracy. We reviewed classical methods and those that incorporate automation to measure the extent of DED [4]. The study showed that DED has yet to benefit from what Artificial Intelligence (AI) has to offer. Using slit-lamp examinations of the ocular surface we aimed to improve the quantification of the Oxford score [5]. Our proposed method uses unsupervised learning to register frames from the examinations to a common coordinate system. By learning the camera motion and depth simultaneously we are able to track the ocular surface in 3-D,

compensate for eye motion and visualise the full eye. The light source attached to the camera is a challenge and a disturbance when learning egomotion. This was solved through semantic segmentation and adding a new supervision signal: semantic reconstruction loss. We also used the advantage of estimating the shape of the eye as prior knowledge we could include as a constraint. This was implemented through a shape fitting loss; the shapes being two spheres intersecting each other. Our registration showed quantitative and qualitative improvement with each contribution. We also calculated the inter-rater reliability of the punctate dots (damaged areas) annotations. Our method came closest to what can be considered human error. The proposed registration method was also used for a pre-processing task, frame selection. Once applied to automated Oxford score classification, our method improved the results as well. The improvement validates that the strong color/illumination variances present in the examinations are a disturbance for any deep learning task. We overcame this in both tasks via our contributions and proposed method.