Deep multimodal visual data fusion for outdoor scenes analysis in challenging weather conditions

Sijie Hu

université PARIS-SACLAY

Deep Multimodal Visual Data Fusion for Outdoor Scenes Analysis in Challenging Weather Conditions

Deep learning pour la fusion multimodale d'images : application à l'analyse de scènes extérieures dans des conditions difficiles

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n°580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat: Sciences du traitement du signal et des images

Graduate school: Sciences de l'Ingénierie et des Systèmes
Référent: Université d'Évry Val d'Essonne

Thèse préparée à l'unité de recherche IBISC (Université Paris-Saclay, Univ Evry), sous la direction de **Désiré SIDIBE** (Professeur), la co-direction de **Samia BOUCHAFA** (Professeur) et le co-encadrement de **Fabien BONARDI** (Maître de Conférences).

Thèse soutenue à Paris-Saclay, le 6 octobre 2023, par

Sijie HU

Composition du Jury
Membres du jury avec voix délibérative

Anissa MOKRAOUI Professeur, Université Sorbonne Paris Nord (L2TI)	Présidente
Sylvie CHAMBON Maître de Conférences - HDR, Toulouse INP (IRIT)	Rapporteur
Yassine RUICHEK Professeur, Université de Technologie Belfort-Montbéliard (CIAD)	Rapporteur
Antoine MANZANERA Professeur, ENSTA Paris (U2IS)	Examinateur

Thèse de doctorat

NNT : 2023UPAST121

Acknowledgements

Firstly, I want to express my heartfelt gratitude to my supervisor, Dr. Prof. Désiré Sidibé, whose wisdom and guidance have constantly helped me advance in my research topic. During every phase of idea realization, he has shared copious insights, thereby providing me with in-depth viewpoints and distinctive wisdom. His substantial assistance in experimental design and manuscript composition has been of enormous value to me. Furthermore, I would like to thank Dr. Prof. Samia Bouchafa and Dr. Fabien Bonardi, my co-supervisors from the IBISC Lab. Their professional knowledge and technical insights have deepened my understanding of the research topic.

At the same time, I would like to express my thanks to my colleagues, Abanob Soliman and Tabia Ahmad. Their support and assistance reaffirmed my belief that the path of scientific research, though tough, is not embarked upon alone. It is worth mentioning that I deeply cherish the times during the IBISC Lab Days. There, colleagues from different research groups shared their research progress and conducted various interesting activities, bringing me countless inspirations and insights, allowing me to deeply feel the openness and warmth of the IBISC Lab family. I would like to express my gratitude to Professor Samia, the head of the IBISC, for her strong support of these activities.

I am very grateful to Professor Désiré for giving me the opportunity to participate in a five-month exchange at the National Institute of Informatics (NII) in Japan as an international intern. This has been a precious experience in my life. I would like to thank my supervisor at NII, Dr. Prof. Helmut Prendinger, whose professional guidance in project planning and advancement, as well as many suggestions on presentation skills, have been greatly rewarding. At the same time, I would like to express my gratitude to my colleagues, Artur Gonçalves, Bastien Rigault, Rafael Prates, and Christian Limberg. They shared various interesting ideas during the daily coffee breaks, providing me with much professional assistance and making my life in Japan full of joy.

While expressing my gratitude to everyone, I cannot forget those who silently support me from behind, my parents. Their support during my pursuit of a Ph.D. allowed me to devote myself fully to research. Their love and trust are the driving forces for my advancement. Furthermore, I would like to express my deepest gratitude to my wife, Dr. Ying Huang. Her support and help in life allowed me to journey down the academic road without any worries. I appreciate her for accompanying me on this journey, which has been both challenging and rewarding.

Résumé en Français

La perception de l'environnement est un aspect critique de la vision par ordinateur, car elle s'efforce d'extraire des caractéristiques et de déduire des informations à partir des entrées, facilitant ainsi une compréhension globale des environs d'un système. Cependant, les environnements du monde réel sont souvent dynamiques et complexes, et un seul capteur est généralement insuffisant pour capturer toutes les informations nécessaires pour les décrire avec précision.

La fusion de données visuelles multimodales, comparée au traitement de données unimodales, peut fournir une compréhension plus complète et précise de l'environnement en intégrant des informations provenant de divers capteurs. Cette stratégie aide à améliorer la perception de l'environnement dans des scénarios dynamiques. Cependant, l'utilisation de données multimodales pose de nouvelles questions :

- Comment pouvons-nous identifier la représentation conjointe optimale pour résoudre la redondance parmi plusieurs entrées ?
- Comment pouvons-nous intégrer et interpréter efficacement les données pour exploiter pleinement la complémentarité inhérente entre différentes modalités ?
- Comment pouvons-nous développer des modèles de fusion efficaces pour satisfaire le besoin de traitement en temps réel ?

Pour répondre à ces questions, plusieurs stratégies de fusion, telles que la fusion précoce, tardive et intermédiaire, ont été étudiées. La fusion précoce, en particulier, fusionne les données d'entrée de plusieurs capteurs au niveau du pixel, permettant au modèle de Deep Learning (DL) d'apprendre des représentations conjointes directement à partir des données fusionnées. En revanche, la fusion tardive traite chaque modalité indépendamment et fusionne les décisions ou prédictions qui en découlent, permettant ainsi au modèle de préserver les informations spécifiques à chaque modalité. La fusion intermédiaire se situe entre les deux, combinant des caractéristiques à différents stades du processus de traitement. Comparées aux approches unimodales, ces stratégies de fusion multimodales ont démontré des performances supérieures dans diverses tâches de vision par ordinateur, s'avérant particulièrement efficaces dans des conditions météorologiques difficiles.

D'autre part, l'adaptation de domaine (DA) vise à transférer les connaissances acquises dans un domaine vers un domaine différent mais connexe, améliorant ainsi les capacités de généralisation des modèles DL. Dans ce scénario, le modèle DL est capable d'exploiter un grand nombre de jeux de données synthétiques correctement annotés, au lieu de compter sur des données largement étiquetées à la main, ce qui réduit considérablement le temps et l'effort nécessaires pour le processus d'annotation des données. De plus, l'utilisation de signaux de supervision provenant de modalités auxiliaires, présente une approche prometteuse pour atténuer les défis associés au décalage de distribution et à l'incertitude des motifs à travers différents domaines.

Par conséquent, cette thèse se concentre sur les problèmes de l'apprentissage multimodal basé sur la modalité de couleur, c'est-à-dire le RGB, en conjonction avec la modalité de profondeur ou d'infrarouge thermique. Plus précisément, nous nous concentrons sur la recherche de stratégies de fusion multimodales efficaces qui exploitent les caractéristiques uniques des différentes modalités pour faire face aux défis posés par les scènes dynamiques, puis sur l'exploration du potentiel des informations multimodales pour bénéficier aux modèles dans l'apprentissage de représentations cohérentes entre les domaines. À cette fin, pendant le développement de cette thèse, nous passons de domaines unimodaux à des domaines multimodaux et explorons des objectifs d'apprentissage plus généralisés. Par la suite, nous avons utilisé des tâches en aval telles que la segmentation sémantique ou la détection d'objets pour évaluer la capacité à interpréter divers scénarios. Les principales contributions sont résumées comme suit :

1. Dans la première contribution, nous nous concentrons sur un cadre populaire de segmentation sémantique connu sous le nom d'encodeur-décodeur et soulignons que les décodeurs existants ne parviennent pas à exploiter de manière exhaustive les informations extraites par l'encodeur. Par conséquent, nous proposons un paradigme composé de deux branches, c'est-à-dire les branches principales et auxiliaires, avec presque aucun paramètre supplémentaire. De plus, nous concevons une stratégie de calcul de fonction de perte en tenant compte des informations de contours. Notre approche permet à différentes branches d'apprendre de manière adaptative des informations complémentaires. Les résultats de nos expériences montrent une amélioration constante des performances des modèles originaux d'encodeur-décodeur sur les scénarios extérieurs, et l'apprentissage d'informations complémentaires peut amener les deux branches à se concurrencer dans une certaine mesure pendant le processus d'apprentissage, ce qui améliore encore les performances (voir Chapitre 3). Les résultats de ce travail ont été présentés lors de la conférence VISAPP 2022 [1].

2. Dans la deuxième contribution, nous ciblons l'analyse de scène multimodale et explorons des stratégies de fusion de données RGB et de profondeur (D). Bien que les méthodes basées sur l'auto-attention aient démontré l'efficacité de la capture des dépendances à longue portée, le coût énorme limite considérablement l'application de cette idée dans la fusion multimodale. À cette fin, nous concevons un bloc de fusion transmodale et sa variante efficace basée sur un mécanisme d'attention additive pour capter efficacement l'information

globale parmi les différentes modalités. Ensuite, nous présentons un bloc de trans-contexte simple mais efficace basé sur un transformeur pour connecter les informations contextuelles. Avec ces conceptions, nous proposons le modèle HCFNet, qui peut explorer les dépendances à longue portée de l'information multimodale tout en conservant les détails locaux. Les expériences montrent que notre mécanisme d'attention aide à former une compréhension globale inter- et intra-modalités. De plus, nos méthodes surpassent les méthodes multimodales actuelles (voir Chapitre 4). Les résultats de ce travail ont été présentés lors de la conférence ICPR 2022 [2].

3. Dans la troisième contribution, ciblant le problème de la détection d'objets dans des conditions de faible luminosité, nous étudions de manière approfondie les stratégies de fusion d'images RGB et thermiques pour améliorer la capacité de perception d'un modèle en utilisant des indices d'imagerie thermique. Pour la fusion RGB-T, nous proposons un module de fusion croisée de patchs (CPCF) pour extraire des caractéristiques multimodales à la fois dans les dimensions spatiales et de canal, au cours duquel le module CPCF exploite de manière adaptative les propriétés spécifiques à une modalité pour calibrer les caractéristiques de l'autre modalité, modélisant ainsi efficacement les propriétés complémentaires entre les modalités et optimisant la représentabilité des caractéristiques dans le flux de données. De plus, nous concevons un cadre de fusion intermédiaire basé sur CPCF, qui peut être intégré de manière flexible dans diverses méthodes de détection d'objets pour exploiter efficacement les indices multimodaux afin d'améliorer les performances des modèles. Les expériences démontrent que notre méthode proposée surpasse d'autres techniques sur une variété de bases de données de référence. De plus, nous montrons qu'il peut être étendu à différents types de détecteurs, illustrant ainsi davantage sa robustesse et son universalité (voir Chapitre 5).

4. Dans la quatrième contribution, nous étudions l'adaptation de domaine non supervisée (UDA) basée sur des données multimodales. Récemment, la profondeur s'est avérée être une propriété pertinente pour fournir des indices géométriques pour améliorer la représentation RGB. Cependant, les méthodes UDA existantes traitent uniquement les images RGB ou exploitent la profondeur avec une tâche auxiliaire d'estimation de la profondeur. Ainsi, nous proposons une nouvelle méthode UDA multimodale nommée MMADT, qui repose sur les images RGB et de profondeur comme entrée pour améliorer la capacité d'adaptation en exploitant les indices géométriques dans la modalité de profondeur. Pour ce faire, nous concevons un simple bloc de fusion de profondeur (DFB) pour recalibrer la profondeur d'entrée et l'aligner avec les caractéristiques RGB. Ensuite, nous alignons explicitement la distribution des caractéristiques de profondeur par un entraînement adversarial. De plus, nous présentons un réseau assistant d'estimation de profondeur multimodal auto-supervisé nommé Geo-Assistant pour transférer l'attention géométrique à notre modèle UDA. Ces stratégies UDA permettent au modèle d'apprendre des représentations plus cohérentes entre les modalités et les domaines. En conséquence, notre méthode améliore considérablement les performances d'adaptation et surpasse les

méthodes basées uniquement sur le RGB (voir Chapitre 6). Les résultats de ce travail ont été publiés dans la revue Pattern recognition [3].

Ce manuscript de thèse est organisé en sept chapitres.

Chapitre 1: Introduction Ce chapitre présente le contexte de la thèse et les pistes de travail envisagées.

Chapitre 2: Etat de l'art Dans ce chapitre, nous avons d'abord passé en revue la progression des techniques de perception de scène basées sur le DL, en mettant particulièrement l'accent sur les tâches impliquant la segmentation sémantique et la détection d'objets. Par la suite, nous avons réfléchi aux stratégies liées à la fusion multimodale, incluant la fusion précoce, tardive et intermédiaire. Cette discussion a principalement englobé les méthodes pour la fusion des images RGB avec des cartes de profondeur, ainsi que la fusion des images RGB avec des images thermiques. Enfin, nous nous sommes plongés dans les travaux liés à l'UDA et avons passé en revue les méthodes basées sur la divergence, l'étiquetage pseudo-label, l'adversarial, et les modalités auxiliaires. Nous avons observé que les approches multimodales pouvaient fournir des informations complémentaires et spécifiques à la scène, renforçant ainsi la stabilité et la précision de la perception de l'environnement. De plus, les techniques liées à la multimodalité peuvent être largement employées dans un ensemble diversifié de tâches pour aider à développer des modèles plus adaptatifs et complets.

Chapitre 3: Une architecture générale de décodeur à deux branches pour la segmentation sémantique Dans ce chapitre, nous nous penchons sur les modèles de segmentation sémantique basés sur les images RGB. L'un de nos objectifs est d'améliorer la sensibilité du modèle aux contours. Pour y parvenir, nous utilisons des signaux de contours sémantiques dans la fonction de coût, qui fournit un signal de supervision supplémentaire pour fournir une segmentation plus précise et efficace. De plus, nous nous concentrons également sur la structure du décodeur et concevons un décodeur générique à deux branches qui pourrait être appliqué de manière flexible aux modèles existants basés sur le modèle encodeur-décodeur et obtenir des gains de performance cohérents. Plus précisément, nous présentons un paradigme général de décodeur à deux branches composé d'une branche principale et d'une branche auxiliaire pour la segmentation de scènes. Ce paradigme de décodeur peut être directement appliqué dans un cadre d'encodeur-décodeur pour affiner et intégrer efficacement l'information extraite par l'encodeur. Avec ce décodeur à deux branches, nous proposons en outre une fonction de perte complémentaire renforcée par l'information de contour appelée BECLoss pour guider les deux branches à apprendre des informations complémentaires. Les expériences comparatives montrent que le paradigme de décodeur à deux branches proposé et BECLoss peuvent améliorer de manière significative les performances du modèle original d'encodeur-décodeur de manière cohérente sur des ensembles de données extérieurs difficiles. De plus, bien que nous ajoutions une

branche au décodeur, cela n'augmente pas de manière significative le nombre de paramètres, et la branche ajoutée peut être supprimée dans le processus d'inférence tout en obtenant des performances bien supérieures à l'original.

Chapitre 4: Un réseau hybride RGB-D Cross Fusion pour la segmentation sémantique Dans ce chapitre, nous approfondissons l'exploration de l'exploitation des images de profondeur pour fournir des indices géométriques supplémentaires dans le cadre d'un framework multimodal. L'objectif est d'améliorer les capacités perceptuelles des modèles de segmentation sémantique, dans lesquels nous utilisons des mécanismes d'attention croisée pour se concentrer sélectivement sur les caractéristiques saillantes d'une modalité tout en minimisant celles qui sont moins informatives. Plus précisément, nous concevons une nouvelle méthode de fusion de données visuelles multimodales, qui peut intégrer efficacement des données provenant de différentes modalités. Elle garantit également que le modèle conserve des détails locaux précieux après la fusion tout en ayant un champ réceptif global. Précisément, nous personnalisons un bloc de fusion multimodal nommé bloc AC basé sur le mécanisme d'attention additive, qui aide à acquérir une information globale inter-modalités et intra-modalités. Ensuite, nous proposons le bloc EAC, une variante efficace du bloc AC, pour construire efficacement une attention globale et conserver des détails dans une entrée haute résolution. D'autre part, sur la base des modèles transformeurs, nous proposons un bloc de fusion de contexte simple mais efficace appelé bloc de contexte (TC) pour connecter davantage la sortie de contexte de l'encodeur. Enfin, avec les composants bien conçus que nous proposons, nous présentons la méthode HCFNet pour la segmentation sémantique des scènes intérieures et extérieures. Des expériences complètes et des études d'ablation vérifient l'efficacité de notre réseau et de ses différents composants.

Chapitre 5: Fusion de données croisées Channel-Patch pour la détection d'objets RGB-T Dans ce chapitre, nous présentons une méthode de fusion multimodale appelée CPCF pour la détection d'objets multispectraux, qui comprend une attention croisée canal-par-canal (CCA), une attention croisée patch-par-patch (PCA) et un module de réglage adaptatif (GA). La CCA et la PCA sont conçues pour affiner les indices précieux provenant des dimensions spatiales et des canaux, respectivement, et exploitent les caractéristiques d'une modalité pour calibrer l'autre modalité, intégrant ainsi mieux l'information de différentes modalités. De plus, nous soutenons que l'information multimodale utile contenue dans les dimensions spatiales et les canaux peut varier pendant le processus de propagation dans le réseau de neurones. Pour tenir compte de cela, nous concevons le module GA pour ajuster adaptativement les poids d'attention dans les dimensions spatiales et des canaux. Par la suite, sur la base de la CPCF, nous concevons une architecture de fusion intermédiaire universelle qui permet une extension à divers types de détecteurs, facilitant l'exploitation de l'information multimodale pour améliorer les performances du modèle. Enfin, nous menons des expériences approfondies avec divers cadres de détection d'objets sur des jeux de données publiques. Les résultats démontrent que notre méthode est capable de capturer efficacement l'information provenant de différentes modalités et de surpasser constamment d'autres méthodes multimodales avancées. De plus, grâce à sa conception

légère, notre méthode peut être intégrée dans des modèles de détection d'objets légers, permettant une détection d'objets en temps réel.

Chapitre 6: Adaptation de domaine multimodale non supervisée Dans ce chapitre, nous présentons un nouveau cadre UDA multimodal pour la segmentation sémantique, qui vise à exploiter des informations supplémentaires pour améliorer les performances d'adaptation. Pour ce faire, nous traitons l'image de profondeur comme une entrée auxiliaire et entraînons le modèle dans un paradigme d'apprentissage multimodal, dans lequel nous rencontrons deux défis. Premièrement, les divergences entre les domaines de la modalité auxiliaire exacerbent encore le fossé entre les domaines. Deuxièmement, les caractéristiques entre différentes modalités ne sont pas nécessairement parfaitement alignées, surtout dans le domaine cible. Pour relever ces défis, nous proposons un réseau multimodal appelé MMADT, composé de trois conceptions clés, à savoir le bloc de fusion de profondeur (DFB), l'entraînement adversarial de profondeur (DAT), et l'assistant Géo (GA). L'application appropriée de ces composants dans un réseau multimodal aide le modèle à atténuer les divergences entre les mêmes modalités dans différents domaines et l'alignement entre différentes modalités dans les mêmes domaines. De plus, l'ajout d'informations spécifiques à la modalité facilite l'apprentissage du modèle UDA d'une représentation de caractéristiques cohérente, améliorant ainsi la capacité de généralisation du modèle sur différents domaines. À notre connaissance, il s'agit du premier travail proposé pour résoudre le problème UDA dans le paradigme d'apprentissage multimodal. De plus, notre stratégie d'entraînement UDA multimodal peut également être librement portée sur les modèles UDA existants. Des expériences approfondies montrent que la modalité supplémentaire peut efficacement améliorer la capacité d'analyse du modèle et résister aux changements de domaine. Nos résultats dépassent largement les bases de référence et les méthodes UDA de l'état de l'art.

Chapitre 7: Conclusion Ce chapitre conclut la thèse par un résumé des principales contributions, et fournit des pistes d'améliorations.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context and Motivation

The pursuit of computer vision is to equip computers and machines with the ability to interpret, analyze, and understand visual information from the surrounding world. Accordingly, environment perception is a critical aspect in computer vision, as it endeavors to extract features and deduce information from inputs, thereby facilitating a comprehensive awareness of a system's surroundings. To this end, several hand-designed operators, such as Sobel [4] and SIFT [5], have been developed to extract boundary information and key features from images. However, these operators, which are tailored to address specific patterns like boundaries or corners, lack efficacy in recognizing and processing more complex features and intricate patterns. Consequently, they fail to adequately address environmental perception's challenge in complex scenarios. In response to these challenges, modern computer vision methods attempt to emulate the way humans garner information from their surroundings. They leverage neural networks to learn intricate patterns and representations from vast amounts of visual data, thereby enhancing the system's capability to extract valuable insights from visual inputs. This progression empowers computer vision systems to interact with and make decisions based on the observed environment, effectively surmounting the challenges posed by conventional image processing methods.

In the past decade, deep learning (DL) techniques based on deep neural networks (DNNs) have found widespread application in the field of computer vision. The emergence and development of these deep models have not only transformed the landscape of computer vision but also enhanced the system's ability to perceive and understand the environment in a broader array of application scenarios, such as faces and emotion recognition, object detection and tracking, and scene reconstruction. Furthermore, to facilitate advancements in DL models and assess their performance effectively, various large-scale datasets have been proposed, such as ImageNet [6], COCO [7], and Pascal VOC [8]. These datasets serve as crucial benchmarks in the field, offering diverse and substantial volumes of data that aid in training models to handle intricate patterns and learning more robust features, which significantly advance

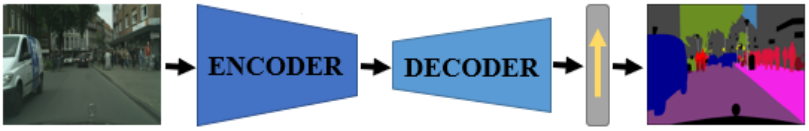


Figure 1.1: General paradigm for semantic segmentation.

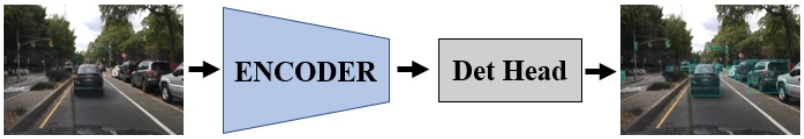


Figure 1.2: General paradigm for object detection.

our ability to perform complex visual tasks. Such tasks include not only basic classification [9] but also higher-level understanding, such as semantic segmentation [10] and object detection [11]. More concretely, semantic segmentation involves assigning a class label to each pixel in an image, as shown in Figure 1.1, while object detection aims to identify and localize specific objects within the scene, as shown in Figure 1.2. By evaluating the performance of DL models on these tasks, researchers can assess their effectiveness in capturing meaningful features and generalizing across various visual scenarios.

On the other hand, breakthroughs in AI algorithms and advancements in hardware technology have led to the widespread application of DL-based visual algorithms in autonomous systems [12]. These systems have achieved considerable success in controlled and ideal conditions, like sunny and well-lit environments. However, such performance tends to diminish when faced with the complexities of real-world scenarios. In fact, real-world environments are frequently open and dynamic, necessitating systems to cope with various challenges stemming from their openness, such as severe performance degradation caused by adverse weather conditions [13].

Recognizing that unimodal data might fall short in providing adequate information for systems to adeptly handle complex and dynamic environments, research has turned towards exploring the potential of multi-modal data. As depicted in Figure 1.3, visible light cameras capture images in the visible spectrum, offering rich color and texture information. However, their performance can diminish in low-light conditions. In contrast, thermal cameras, which are designed to detect infrared radiation, can reveal temperature-based data and excel in dark conditions, yet struggle when tasked with differentiating objects with similar thermal profiles. Furthermore, depth cameras, while contributing valuable distance and geometric information, often underperform in detailing background information or identifying small objects situated at long distances. Consequently, recent scholarly pursuits have sought to utilize multi-modal visual data fusion techniques, aiming to bolster the robustness and adaptability of DL models when faced with diverse and challenging conditions.

Specifically, multi-modal visual data fusion can offer a more comprehensive and accurate understanding of the environment by integrating information from different sensors. For instance, color and texture information from visible light cameras can be combined with temperature data from thermal imagers to better identify and track

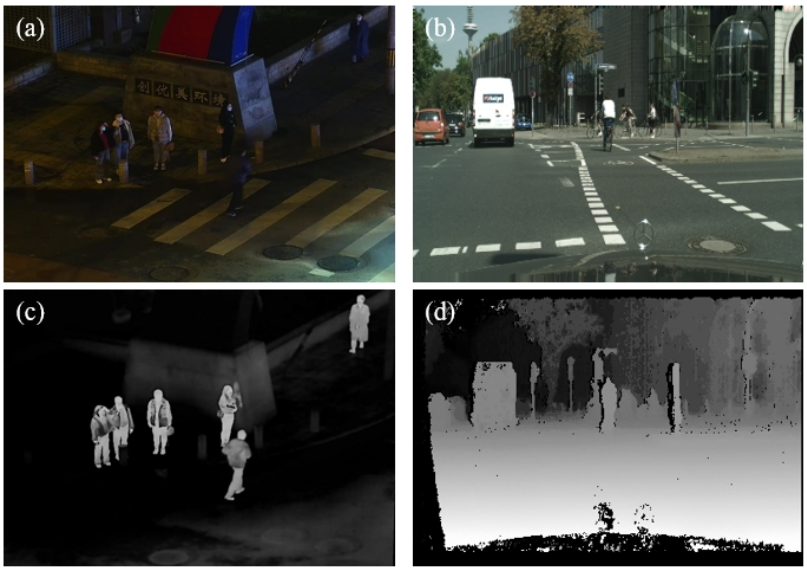


Figure 1.3: Image captured in different cameras. (a), (b) were captured by visible light cameras, (c) was captured by thermal camera and (d) was produced by stereo camera (depth).

objects in adverse weather conditions. Simultaneously, distance and geometric information from depth cameras can be used to improve object localization and navigation. This multi-sensor approach leverages the unique capabilities of each device to provide a more robust and nuanced perception of the environment. Therefore, multi-modal-based DL models can better handle the complexities and uncertainties associated with real-world scenarios, allowing for improved performance in various conditions. Nonetheless, utilizing multi-modal data introduces new questions: How can we identify the optimal joint representation to address the redundancy among multiple inputs? How can we effectively integrate and interpret data to fully exploit the inherent complementarity between different modalities? Additionally, how can we develop efficient fusion models to satisfy the need for real-time processing?

To answer these questions, several fusion strategies, such as early, late, and intermediate fusion, have been investigated by researchers [14]. Early fusion, specifically, merges input data from multiple sensors at the pixel level, enabling the DL model to learn joint representations directly from the fused data. In contrast, late fusion treats each modality independently and fuses the ensuing decisions or predictions, thereby permitting the model to preserve modality-specific information. Intermediate fusion falls in between, combining features at different stages of the processing pipeline. Thereby, compared with unimodal approaches, these multi-modal fusion strategies have demonstrated superior performance in various computer vision tasks, proving particularly in challenging weather conditions.

On the other hand, domain adaptation (DA), as a branch of transfer learning, aims to transfer knowledge learned from one domain to a different but related domain, enhancing the generalization capabilities of DL models [15]. In contrast to standard deep learning techniques, DA focuses on enabling models to learn general representations across domains, thereby facilitating the acquisition of consistent information from outdoor scenarios under dynamic

conditions. Furthermore, another motivation behind domain adaptation is to deploy models trained on a specific domain with annotated data in new domains where annotations are limited or unavailable. In this scenario, the DL model is able to leverage a large number of accurately annotated synthetic datasets, such as GTA5 [16] and SYNTHIA [17], instead of relying on extensive hand-labeled data, which, in turn, significantly reduces the time and effort required for the data annotation process. In addition, utilizing supervisory signals from auxiliary modalities in DA presents a promising approach to alleviate the challenges associated with distribution shift and pattern uncertainty across different domains [18, 19, 20]. These auxiliary modalities, which can function as guiding signals during the adaptation process, offer essential contextual clues that aid in achieving more precise feature alignment between the domains.

Motivated by these observations, this thesis seeks to investigate efficient multi-modal visual data fusion strategies and delve into leveraging multi-modal information to enhance the model's generalization capabilities across different domains, ensuring a robust performance of DL models in outdoor scene understanding scenarios under challenging weather conditions.

1.2 Background

The concept of multi-modal visual data fusion has been extensively explored in semantic segmentation and object detection tasks, which involves the integration of homogeneous data inputs - such as RGB images, infrared images, and depth maps - into a single analysis framework. Technically, the term "modality" refers to each sensor or detector capturing information about the same scene [21]. For instance, in robotics, primary modalities commonly encompass devices such as color cameras, near-infrared cameras, depth cameras, and Event cameras [22]. Meanwhile, in autonomous driving, available sensors also include LiDAR and Radar. Moreover, audio and text modalities in multimedia applications provide valuable information specific to certain scenarios [23]. In this context, the fusion process is intended to harness the unique strengths and perspectives provided by each data modality, thereby providing a more comprehensive understanding of the scene or object under investigation. Typically, for semantic segmentation, deep multi-modal fusion methods strive to automatically learn optimal semantic mappings, where they take advantage of the complementary information derived from the same scene [24, 17, 25]. As for object detection, these methods employ multi-modal data to enhance the model's ability to identify and locate objects of interest, while simultaneously reducing its sensitivity to the background [26, 27].

1.3 Open Research Questions

1.3.1 Multi-modal Visual Data Fusion

Multi-modal visual data fusion targets integrating information from various visual sensors to form a comprehensive understanding of a given scenario. Generally, each modality contributes some additional information to the overall system, which helps to constrain the degrees of freedom for environmental variables within dynamic scenes [28]. Moreover, it offers diverse perspectives for a unified scene, which provides supplementary cues to reveal specific patterns within the given observations. Therefore, learning from multiple modalities is beneficial. However, different modalities represent data types in distinct sampling spaces, such as color images are captured by measuring the intensity of light reflected off objects within the visible spectrum, whereas thermal infrared images sample the thermal infrared radiation emitted from the surface of objects [29]. Although the primary intention of using multi-modal data is to leverage the complementarity of different modalities, the information redundancy that comes with multi-modality can also pose challenges in learning optimal feature representation of multi-modal data fusion [30].

In this context, the main challenges involve multi-modal fusion, feature alignment, and collaborative learning. To be more specific, multi-modality employs various data forms for depicting the same scene, and thus extracting generalized feature expressions from mixed representations is crucial in enhancing the model's perceptual capabilities. Then, the primary concern during the fusion process is extracting valuable information from different modalities and integrating it into a unified multi-modal representation. On the other hand, feature alignment efforts to model the correspondence between different modalities, which encompass both semantic and spatial alignment. Commonly, the spatial alignment of visual modalities can be simplified through the calibration of sensors, whereas semantic alignment largely depends on the model's knowledge to extract details from different modalities. Furthermore, since the alignment is not explicitly specified in the data, there may not exist a one-to-one correspondence between features of different modalities. For instance, due to limitations in detection range, some objects present in color images might not have corresponding representations in depth images. Collaborative learning poses its own difficulties, as it involves the joint learning of the underlying interdependencies between modalities, which typically requires models to learn how to share and transfer knowledge across modalities. Furthermore, leveraging knowledge transfer between modalities to mitigate the impact of modality noise and missing modalities also presents challenges.

1.3.2 Domain Adaptation

Domain adaptation (DA) is a technique employed to learn consistent and transferable feature representations by incorporating DA methods into traditional DL pipelines [15]. In many real-world applications, the majority of supervised learning assumes that training and testing data come from the same distribution. However, this is not

always the case, and thus the performance of the model might experience a significant decline. As a particular case of transfer learning, DA is designed to address situations where a model is trained on one or more labeled source domains and is expected to perform competitively when working on a related but distinct target domain. This is especially applicable for handling scenarios in which there are substantial differences between the training and testing environments due to factors such as weather conditions in practical applications.

However, DA also introduces challenges stemming from data distribution mismatches, label discrepancies, and data imbalance. In more specific terms, distribution mismatches can be attributed to various factors. For instance, differences in data acquisition devices and environments inevitably lead to discrepancies in the feature distributions between the source and target domains. Moreover, underlying patterns within the data might also exhibit differences, such as data in the source domain is generated by a simulator [17]. Although the simulated data is derived from real-world scenarios, substantial differences still exist in higher-order features, such as the specific presentation of instances and the style of scenes. On the other hand, due to differences in task types or annotation strategies, the label spaces of the source and target domains are only partially congruent. For instance, in object detection tasks, the labels in the source domain may encompass bounding boxes for multiple classes of objects, whereas the target domain may focus exclusively on detecting a specific type of object. This discrepancy leads to inconsistent feature transfer and feature redundancy, which can ultimately impair the model's adaptability. Additionally, another challenge in DA is data imbalance, which is quite prevalent in DL since collecting and annotating large amounts of high-quality data can be both difficult and costly, e.g., the annotation of each real image in the Cityscapes dataset requires approximately 1.5 hours for semantic segmentation tasks [31]. In this context, the source domain is characterized by an abundance of labeled data, whereas the target domain may have only a limited amount of labeled data or none at all, which we refer to the former case as semi-supervised DA [32] and the latter as unsupervised DA [33]. Besides, the complexities of training strategies and multi-source domain adaptation add another layer to the challenges.

1.4 Contributions

As we previously mentioned, enhancing the perceptual capabilities of a model through multi-modal visual data fusion involves a series of challenges and applications. In this thesis, we mainly focus on the issues of multi-modal learning based on regular color modality, i.e., RGB, in conjunction with depth or thermal infrared modality. Our primary objective is to leverage multi-modal visual data to augment the perceptive capacity of DL models in outdoor scenarios. More specifically, we concentrate on researching effective multi-modal fusion strategies that exploit the unique characteristics of different modalities to deal with the challenges posed by dynamic scenes and then exploring the potential of multi-modal information to benefit models in learning consistent cross-domain representations.

To this end, during the development of this thesis, we transition from uni-modal to multi-modal domains and

explore more generalized DL objectives. Subsequently, we employed downstream tasks such as semantic segmentation or object detection to evaluate the capability to interpret various scenarios. The major contributions are outlined as follows:

1. In the first contribution, we focus on a popular semantic segmentation framework known as encoder-decoder and point out that existing decoders fail to parse the information extracted by the encoder comprehensively. Therefore, we propose a two-branch paradigm composed of two branches, i.e., main and auxiliary branches, with almost no additional parameters. In addition, we design a boundary-enhanced loss computation strategy. Our designs allow different branches to learn complementary information adaptively instead of explicitly indicating the specific learning element. The results of our experiments show that these designs improved the performance of the original encoder-decoder models consistently on outdoor scenarios, and learning complementary information can make the two branches compete with each other to a certain extent during the learning process, which further improves performance (see Chapter 3). The results of this work were presented in the conference VISAPP 2022 [1].

2. In the second contribution, we target multi-modal scene parsing and explore multi-modal cross-fusion strategies based on RGB-D. Although self-attention-based methods have demonstrated the effectiveness of capturing long-range dependencies, the tremendous cost dramatically limits the application of this idea in multi-modal fusion. To this end, we design a multi-modal cross-fusion block and its efficient variant based on an additive attention mechanism to efficiently capture global awareness among different modalities. Then, we present a simple yet efficient transformer-based trans-context block to connect the contextual information. With these designs, we propose light HCFNet, which can explore long-range dependencies of multi-modal information while keeping local details. The experiments show that our attention mechanism assisted in forming global awareness inter- and inner-modalities. In addition, our methods outperformed current multi-modal methods (see Chapter 4). The results of this work were presented in the conference ICPR 2022 [2].

3. In the third contribution, targeting the problem of object detection under low light conditions, we extensively investigate RGB and thermal image fusion strategies to enhance the perception capability of a model to its surroundings by using thermal imaging cues. For RGB-T fusion, we propose a lightweight channel-patch cross fusion (CPCF) module to construct cross-modal features in both channel and spatial dimensions, during which the CPCF module adaptively leverages the properties specific to one modality to calibrate the features of another, thus effectively modeling the complementary properties between modalities and optimizing the representability of features in the data stream. Furthermore, we design an intermediate fusion framework based on CPCF, which can be flexibly integrated into various object detection frameworks to efficiently exploit multi-modal cues to boost the performance of models. Experiments demonstrate that our proposed method outperforms other techniques in a variety of multi-modal benchmarks. Besides, we show that it can be extended

to different types of detectors, thereby further illustrating its robustness and universality (see Chapter 5).

4. In the fourth contribution, we investigate multi-modal-based unsupervised domain adaptation (UDA). Recently, depth has proven to be a relevant property for providing geometric cues to enhance the RGB representation. However, existing UDA methods solely process RGB images or additionally cultivate depth-awareness with an auxiliary depth estimation task. Thus, we propose a novel multi-modal UDA method named MMADT, which relies on both RGB and depth images as input to improve the adaptive capability by leveraging geometric cues in depth modality. To do so, we design a simple Depth Fusion Block (DFB) to recalibrate the input depth and align it with the RGB features. Then, we explicitly align the feature distribution of depth by Depth Adversarial Training (DAT). In addition, we present a self-supervised multi-modal depth estimation assistant network named Geo-Assistant to transfer the geometric attention to our UDA model. These UDA strategies enable the model to learn more consistent representations across modalities and domains. As a result, our method significantly improves adaptation performance and performs favorably against RGB-only-based methods (see Chapter 6). The results of this work are presented in a journal paper [3].

1.4.1 Publications

The thesis main contributions have been published in various scientific papers as given in the following list:

- **Journal papers**

  1. Sijie Hu, Fabien Bonardi, Samia Bouchafa, Désiré Sidibé, "*Multi-modal unsupervised domain adaptation for semantic image segmentation*", **Pattern Recognition**, 137, 109299, 2023.
  2. Sijie Hu, Fabien Bonardi, Samia Bouchafa, Désiré Sidibé, "*Rethinking Self-Attention for Multispectral Object Detection*", **IEEE Trans. on Intelligent Transportation Systems**, Under submission, 2023

- **International conferences**

  1. Sijie Hu, Fabien Bonardi, Samia Bouchafa, Désiré Sidibé, "*A hybrid multi-modal visual data cross fusion network for indoor and outdoor scene segmentation*", **ICPR**, 2022.
  2. Sijie Hu, Fabien Bonardi, Samia Bouchafa, Désiré Sidibé, "*A General Two-branch Decoder Architecture for Improving Encoder-decoder Image Segmentation Models*", **VISAPP**, 2022.

1.5 Organization

This thesis is organized as follows.

- Chapter 2 encompasses an extensive literature review. Among them, we first present the classic works and advancements in the field of deep learning with regard to semantic segmentation and object detection tasks. Subsequently, we delve into the fundamental forms and strategies of multi-modal visual data fusion techniques, discussing the associated works in depth. Lastly, we explore techniques related to unsupervised domain adaptation, with a particular focus on the application of multi-modal data within this context.

- Chapter 3 primarily focuses on RGB-based semantic segmentation methods and proposes an improved two-branch decoder paradigm grounded in the encoder-decoder framework, along with a novel loss function, which aims to improve the training efficiency and segmentation accuracy of the models.

- Chapter 4 explores how to optimize the feature representation of RGB-D data using fusion strategies to enhance the accuracy and efficiency of semantic segmentation models. A multi-modal cross-fusion module, based on the additive attention mechanism, is proposed to capture global awareness between different modalities while preserving local details.

- Chapter 5 investigates the data fusion strategies based on RGB-T data to enhance the model's environmental perception under low-light conditions and proposes a lightweight cross-fusion module to facilitate inter-modal rectification between different modes.

- Chapter 6 focuses on exploring how to leverage multi-modal information to enhance the transfer performance of models in a UDA setting and proposes an RGB-D-based multi-modal UDA method to enhance the geometric cues for semantic segmentation.

- Chapter 7 concludes with the developments and discoveries of this thesis and suggestions for future researches.

Chapter 2

Literature Review

DL-based multi-modal visual data fusion is a broad topic, extensively investigated in various task contexts, encompassing a range of deep learning techniques. It involves the integration of multiple sources of visual data, each providing unique and complementary information, to enhance the learning and decision-making capabilities of the model. In this chapter, we initially introduce two popular tasks within the realm of deep learning-based scene understanding: semantic segmentation and object detection. We briefly review the basic approaches and distinctive works addressing these tasks. Following, we systematically introduce the fundamental frameworks for multi-modal visual data fusion based on the position where the data is integrated within deep models. These positions include early fusion, late fusion, and intermediate fusion. In this context, we review typical works in the application of these methods to the tasks of semantic segmentation and object detection. Finally, we briefly review pertinent works related to unsupervised domain adaptation (UDA) to address the challenges associated with the performance transfer of models across domains.

2.1 Semantic Segmentation

Semantic segmentation is a process where each pixel in the image is categorized into one of the predefined classes, providing a comprehensive understanding of the scene's composition. It is particularly valuable in many real-world applications, such as autonomous driving [34], medical image diagnosis [35], and defect detection [36]. Technically, DL-based semantic segmentation involves training a deep neural network to establish a semantic correspondence between semantic labels and dynamic scene images [37]. As depicted in Figure 2.1, the foundational structure of semantic segmentation entails an encoder-decoder [38] architecture. The encoder, responsible for abstracting high-dimensional features from input images, works in tandem with the decoder, which reconstructs the abstracted features back to the original input resolution while retaining high-level semantic information.

Within this framework, a series of advanced semantic segmentation methods have been developed. On the one

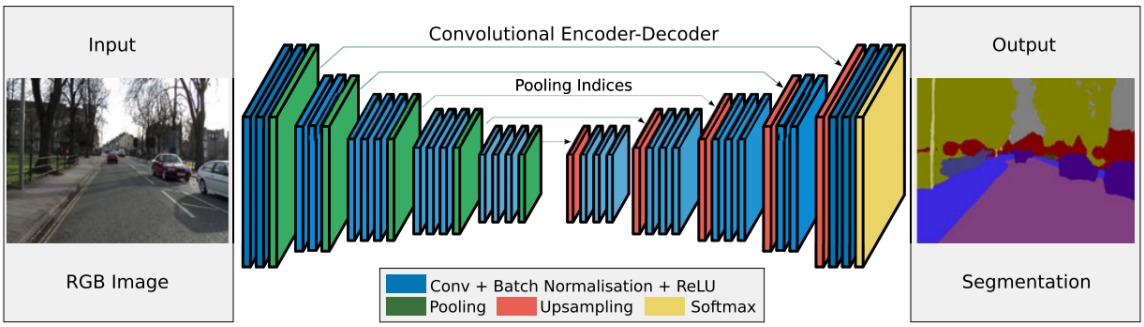


Figure 2.1: Illustration of the encoder-decoder-based SegNet architecture.
The image is from [38]

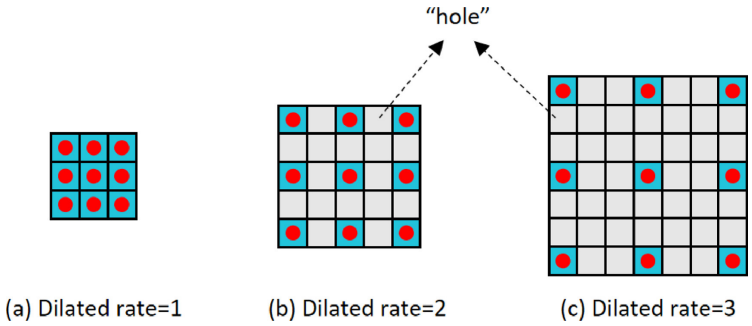


Figure 2.2: Illustration of the dilated convolution kernels with different dilated rates.
The image is from [40]

hand, these methods exploit well-designed backbones, such as VGG [39] and ResNet [9], as feature extractors to encode meaningful information. In addition, DilatedNet [40] employs dilated convolutions for dense predictions, allowing the model to maintain image resolution while capturing multi-scale contextual information, thereby endowing the extracted features with an expanded receptive field. Figure 2.2 illustrate the dilated convolution kernels principle with different dilated rates. Moreover, PSPNet [41] incorporates a pyramid pooling module that aggregates contextual information from varying regions, enhancing its capacity to understand both the global context and local details in a scene. Then, DeeplabV3 [42] integrates atrous convolutions with pyramid pooling module and introduces a module known as the atrous spatial pyramid pooling, which empowers the model to efficiently discern finer details while maintaining a broader contextual understanding.

On the other hand, additional strategies and techniques, such as attention mechanisms [43] and skip-connections [44], are employed to further augment both the performance and efficiency of the segmentation process. For instance, PAN [45] merges the attention mechanism with spatial pyramid structure to extract superior pixel-level feature representations and incorporates a global attention upsampling module at decoder layers to further refine detailed features. Attention U-Net [46] employs an attention gate module to concentrate on task-relevant salient features and inhibit irrelevant areas in the input image, enhancing model focus and performance. Moreover, RefineNet [47] effectively integrates information obtained during the downsampling process via skip connections, enabling high-resolution predictions. Similarly, DeepLabV3Plus [48] incorporate shallow layer information from the

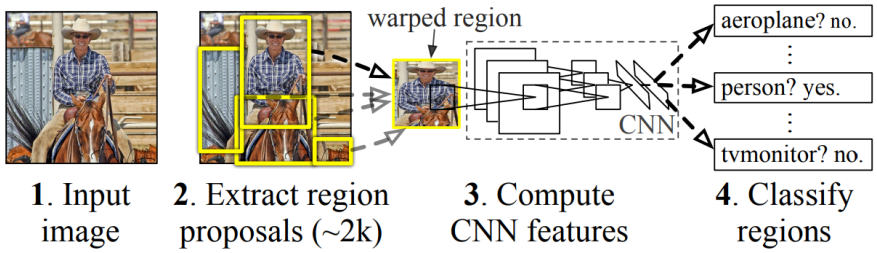


Figure 2.3: Illustration of the two-stage-based object detection system. The image is from [50]

encoder into the decoder to improve fine-grained features. Besides, SegFix [49] proposes a model-agnostic post-processing solution that improves the quality of segmentation boundaries by learning to establish a relationship between boundary pixels and internal pixels.

2.2 Object Detection

Object detection is a prevalent topic in scene perception, which entails identifying and localizing multiple objects within an image. Typically, object detection models can be grouped into two-stage and single-stage based methods. As shown in Figure 2.3, a two-stage-based method divide the detection process into two distinct phases: 1) the region proposal phase, where potential regions of interest within the image are identified, and 2) object classification and bounding box regression phase, where the proposed regions are further abstracted and classified into specific categories and predicting the bounding boxes to encompass the identified objects. As a trailblazing effort, RCNN [50] leverages the selective search algorithm [51] to generate numerous potential regions, then employs SVM and a regressor for classification and bounding box prediction tasks, respectively. After that, Fast-RCNN [11] enhances efficiency by directly extracting the features of the entire image and mapping each candidate region to the extracted feature map, thereby sidestepping the substantial computational burden of individually extracting features for each candidate region, and utilizes a multi-task loss to simplify the multi-stage training process in R-CNN. Faster-R-CNN [52] implements a CNN-based region proposal network, bypassing the need for a selective search algorithm, which enhances the inference speed. Furthermore, FPN [53] introduces a top-down architecture with lateral connections to foster high-level semantics at all scales, thereby significantly enhancing object detection across various scales. While two-stage-based methods demonstrate high performance, their computational complexity and potential slowdown due to multi-stage processing can limit their suitability for real-time or low-latency applications.

Single-stage-based methods endeavor to retrieve all objects in one-step inference. As illustrated in Figure 2.4, the first one-stage-based work is YOLO [54], which applies a single neural network to predict bounding boxes and class probabilities for each grid in an image simultaneously, thereby achieving real-time detection with a reasonable trade-off between speed and accuracy. Following this, a succession of subsequent studies [55, 56, 57] have put

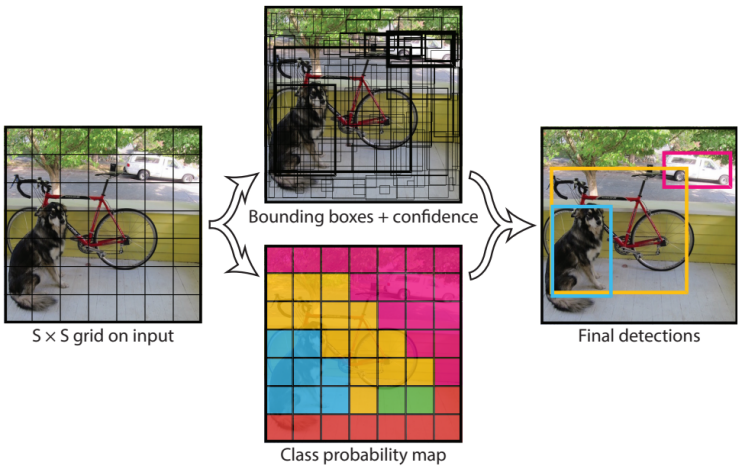


Figure 2.4: Illustration of the one-stage-based object detection system.
The image is from [54]

forward to further improve performance by leveraging various tricks, such as feature pyramid architecture and data augmentation, to strike a balance between detection accuracy and operational speed. Besides, SSD [58] explicitly combines multiple feature maps at different resolutions and produces predictions at different scales, effectively improving the detection performance on small targets. Moreover, RetinaNet [59] targets the problem of foreground-background class imbalance that arises during the training and introduces a novel loss function, i.e., focal loss, which puts more emphasis on hard-to-classify instances. Recently, YOLOv7 [60] incorporates an efficient model structure and introduces a dynamic label assignment strategy, thereby elevating the performance of single-stage models.

Contrary to detectors that depend on hand-designed anchors, recent studies have turned to exploring anchor-free-based methods. This shift has been driven by the recognition that the hand-designed anchor set, which introduces additional degrees of freedom, consequently increases the complexity of model learning. To this end, CornerNet [61] presents a novel approach to representing bounding boxes as a pair of keypoints, namely, the top-left and bottom-right corners, which avoids setting the anchor set manually. Then, FCOS [62] eliminates the predefined set of anchor boxes and directly outputs multi-scale, pixel-level feature abstractions for classifying and bounding boxes regressing, thereby successfully sidestepping the complex computations typically associated with anchor boxes. Recently, YOLOX [63] transforms the YOLO detector into an anchor-free style, which results in enhanced processing speed, and then it capitalizes on strong data augmentation and advanced label assignment strategies to attain superior performance.

Nevertheless, while effective in many scenarios, standard object detection techniques have limitations in handling objects that are not aligned horizontally, as they typically rely on axis-aligned bounding boxes. Figure 2.5 shows that the dense oriented objects fail to be represented with the standard horizontal bounding boxes. To address this challenge, a line of research has been devoted to oriented object detection, where the aim is to predict tight bounding boxes that align closely with the actual orientation of the objects. DRBox [65] presents a novel rotatable bounding

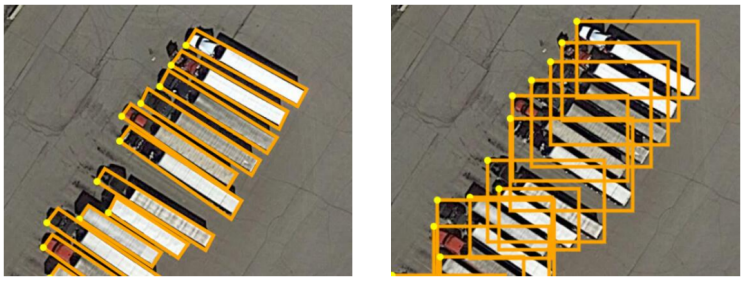


Figure 2.5: Illustration of annotations with oriented bounding boxes and corresponding failure cases with horizontal rectangle annotations.

The image is from [64]

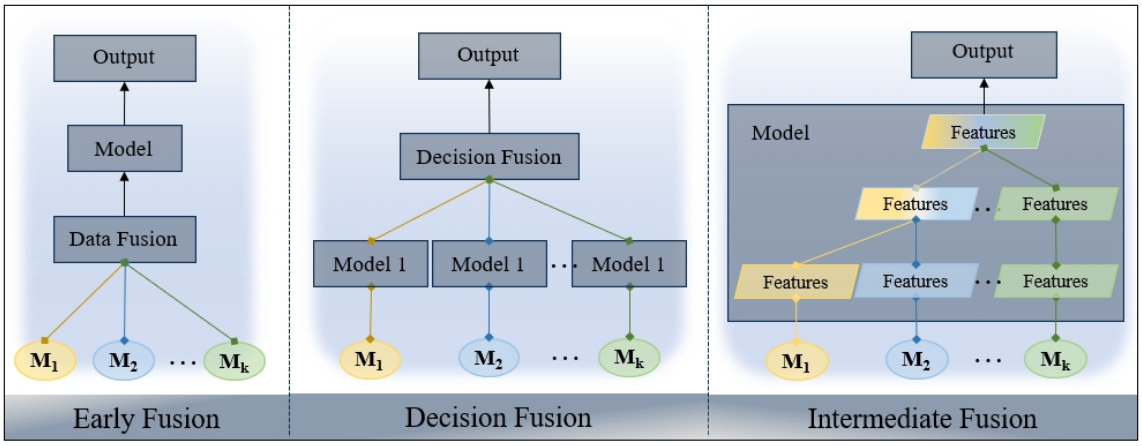


Figure 2.6: Illustration of different multi-modal fusion strategies.

box, which extends the traditional bounding box by incorporating an extra angle parameter, and then predefines a set of anchor sets at different angles for fitting the rotated targets. Moreover, S^2A-Net [66] introduces a feature alignment module and an oriented detection module for mitigating the misalignment between oriented anchors and axis-aligned convolutional features. DRN [67] introduces a feature selection module, which empowers neurons to adapt their receptive fields based on the present state and orientation of the detected object, thus providing robust features to the detection head.

2.3 Deep Multi-modal Visual Data Fusion

Multi-modal visual data fusion is an approach that capitalizes on information from diverse visual modalities to enhance the robustness and accuracy of DL models. This fusion process can be classified into early fusion, late fusion, and intermediate fusion, depending on where the fusion takes place in the model architecture [68]. As illustrated in Figure 2.6, DL architectures offer the flexibility of implementing multi-modal fusion.

Early fusion incorporates multiple modality input features into a single model to learn a unified representation. In situations where multi-modal inputs are spatially aligned, the most straightforward approach is to concatenate data

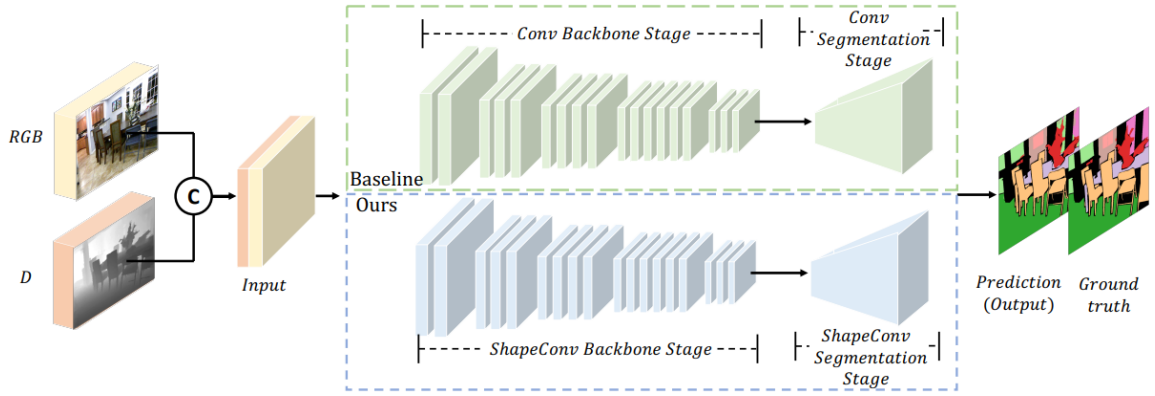


Figure 2.7: Illustration of the ShapeConv-based semantic segmentation network architecture. The image is from [70]

from various modalities into a unified input prior to end-to-end training or utilize learnable parameters to encode multi-modal data and align them at the low-level features. As an initial attempt at multi-modal fusion, early fusion can be simply described as:

$$p = \mathbf{f}([m_1, \ldots, m_n]), \tag{2.1}$$

where $\mathbf{f}$ denotes the DL model, m_i denotes a specific modality and $[.]$ refers to the concatenation operation.

In the late fusion approach, also referred to as decision-level fusion, each modality is processed independently through separate models and the results are fused with a fusion mechanism, such as averaging, voting, or a learned model, at the decision level [69]. We denote $\mathbf{f_i}$ as the model to process a specific modality m_i, $\mathbf{F}$ as a fusion mechanism to merge the output of each model. Then, the final prediction of late fusion can be formulated as:

$$p = \mathbf{F}(\mathbf{f_1}(m_1), \ldots, \mathbf{f_n}(m_n)). \tag{2.2}$$

During this process, features from various modalities are handled independently, which consequently leads to inadequate modeling of intermodal features interactions.

On the other hand, leveraging the flexibility inherent in DL architectures, intermediate fusion was proposed to integrate the advantages of both early and late fusion strategies, in which features from different modalities interact with each other while still preserving their individuality. The intermediate fusion-based model can be described as:

$$p = \mathbf{f}(\mathbf{g_1}(m_1), \ldots, \mathbf{g_n}(m_n)), \tag{2.3}$$

where g_i denotes the sub-network used to process modality m_i. Furthermore, intermediate fusion strategies are often employed in conjunction with a variety of deep learning techniques, such as attention mechanisms [43], to create a rich and versatile feature space where complementary information or noise in different modalities are adaptively highlighted or attenuated.

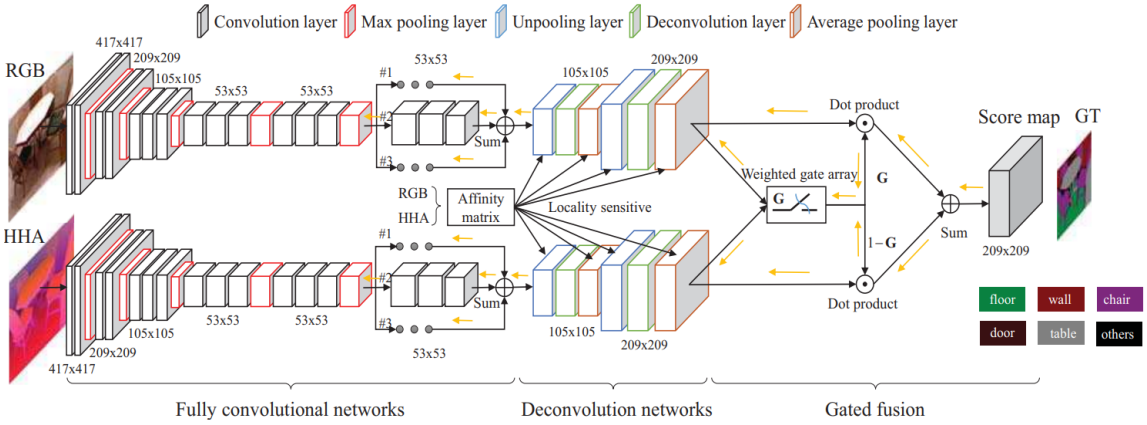


Figure 2.8: Illustration of the late fusion-based LSD-GF model for semantic segmentation. The image is from [73]

2.3.1 Early Fusion

As an initial exploration of early fusion, [71] combines RGB and depth maps into a four-channel input and employs a multiscale convolutional network for multi-modal feature extraction, which, while promising, fails to fully exploit the potential of multi-modal fusion due to inherent limitations of the original DL network. For instance, previous work [72] has revealed that fusing different modalities at the pixel level overlook features specific to each modality, thereby compromising the accuracy of detection. Moreover, [72] applies a strategy that concatenates RGB and thermal images after processing them through one convolutional layer and operates them in conjunction with a pre-trained backbone network for feature extraction. The aim is to leverage early fusion techniques to capture lower-level visual features such as corners and line segments. Besides, ShapeConv [70] is a recently proposed method that employs a model-agnostic shape-aware convolutional layer to decompose features of depth maps and then uses reparameterization techniques to reorganize learned weights into the standard convolution operation during inference, as illustrated in Figure 2.7.

2.3.2 Late Fusion

To extract modality-specific features from the input, the LFC [75] first trains segmentation expert networks for different modalities and perform element-wise summation of the feature maps produced by these networks, followed by a series of convolution, pooling, and up-convolution operated to tune the final output. Subsequently, CMoDE [76] improves the architecture of expert networks and designs class-specific gate networks and fused convolutions, and then during training, it freezes the parameters of expert networks to compel the gate network to utilize representations learned by the experts, thereby leveraging the complementary features of the experts. Similarly, LSD-GF [73] employs a learnable gating network to automatically learn the varying contributions of each modality in different scenes for classifying different categories, thereby more effectively combining RGB and depth cues, as shown in Figure 2.8. In addition, IAF-R-CNN [77] utilizes two separate subnetworks to process RGB and thermal images,

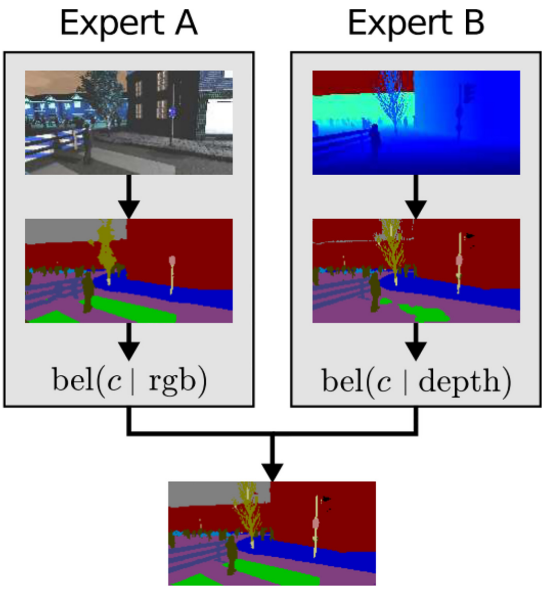


Figure 2.9: Illustration of the statistical fusion methods for combining different experts. The image is from [74]

respectively, with an additional branch network to estimate fusion weights for an optimized combination of outputs from both subnetworks. Besides, [74] proposes a statistical-based post-processing method that utilizes Bayesian or Dirichlet fusion for statistical merging, enabling individual experts to be trained on different datasets without the need for additional training to integrate their outputs, as shown in Figure 2.9.

2.3.3 Intermediate Fusion

A representative work is FuseNet [78], as illustrated in Figure 2.10. It employs an auxiliary encoder to extract depth-related features, which are then fused with color information and passed through a decoder network to restore semantic information. Building upon the foundations of FuseNet, the LDFNet [79] advances the structure of the auxiliary encoder and introduces luminance information to calibrate depth images. Moreover, RFBNet [80] proposes residual fusion block, which serves to model the interdependencies among various encoders, thereby progressively aggregating modality-specific features and cross-modal features from these encoders. In addition, [81] proposes a fusion mechanism known as self-supervised model adaptation (SSMA) to establish correlations between two modality-specific feature maps, which enables the network to emphasize more informative features selectively within one modality while suppressing less informative features in another. Then, a skip-connection is employed to integrate multi-scale fused features into the decoder, as shown in Figure 2.11.

More recently, a series of multi-modal fusion strategies based on the attention mechanism have been proposed to steer the alignment of cross-modal features. For instance, ACNet [82] introduces an attention complementary module that extracts weighted features from RGB and depth branches based on channel attention and utilizes an additional encoder structure to handle the fused features. Sharing a similar idea, [83] develops the separation-

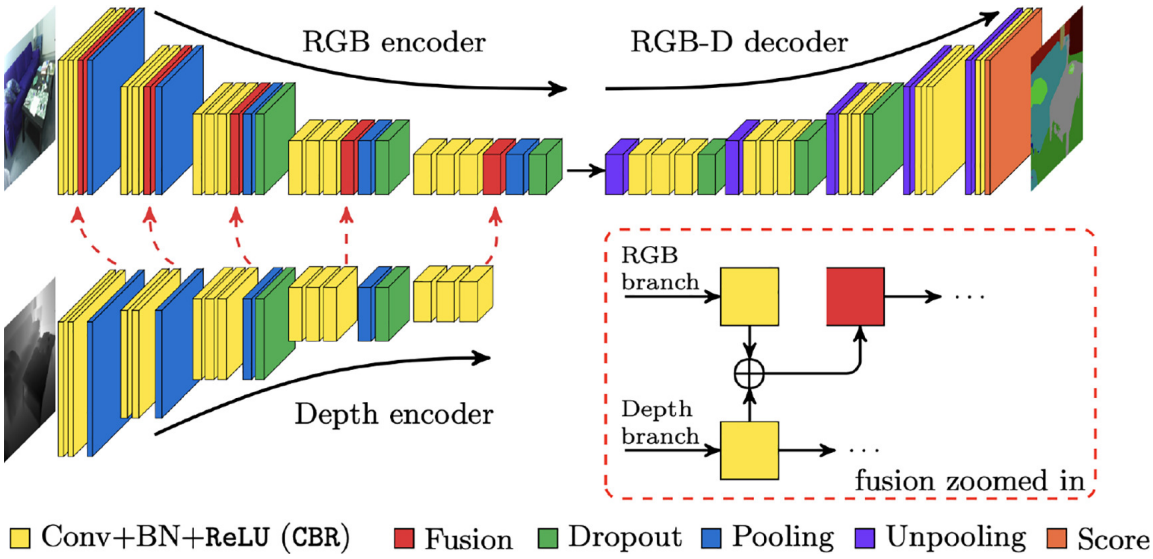


Figure 2.10: Illustration of FuseNet architecture with RGB-D input.
The image is from [78]

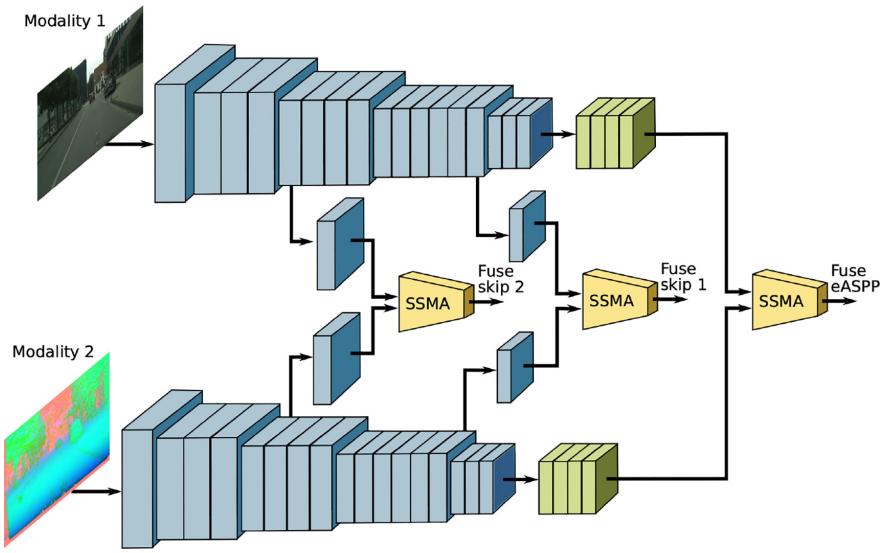


Figure 2.11: Illustration of fusion architecture with SSMA block.
The image is from [81]

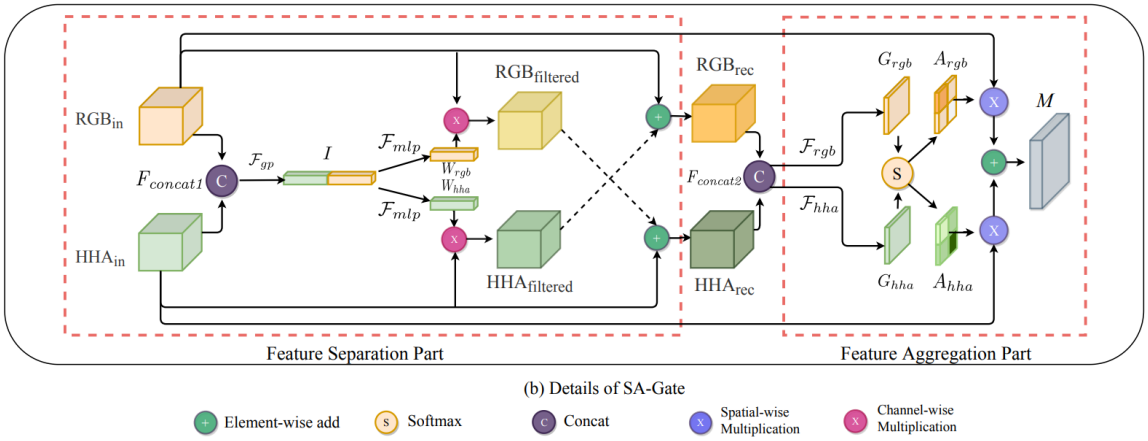


Figure 2.12: Illustration of SA-Gate.
The image is from [83]

and-aggregation gate (SA-Gate) to suppress depth data noise and recalibrate corresponding RGB features, and then through a gated module to blend the cross-modal information. The fusion pipeline is shown in Figure 2.12. Alternatively, AFNet [84] incorporates a co-attention mechanism into an attention fusion module, enhancing the contextual correlation between RGB and IR feature maps by considering their cross-spectral complementarity at the final stage of the encoder. Recently, CMAFF [85] developed a lightweight multispectral feature fusion technique to infer shared and differential information from the intermediate feature maps of RGB and thermal IR modalities, which is then used to calibrate multi-modal features.

2.4 Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) is a branch of transfer learning that concentrates on situations where source data is labeled during training, but the target data is not. In practice, however, a domain discrepancy often exists between source and target domains. This discrepancy can lead to less-than-optimal performance in model transferability. Formally, in UDA setup, given a source domain $\mathcal{D}_{\mathcal{S}} = \{\mathcal{X}_{\mathcal{S}}, \mathcal{X}_{\mathcal{S}}\}_{i=1}^{\mathcal{N}_{\mathcal{S}}}$ with $\mathcal{N}_{\mathcal{S}}$ labeled samples and a target domain $\mathcal{D}_{\mathcal{T}} = \{\mathcal{X}_{\mathcal{T}}\}_{i=1}^{\mathcal{N}_{\mathcal{T}}}$ with $\mathcal{N}_{\mathcal{T}}$ samples without any labels, the primary objective is to minimize the domain gap between the labeled source data and unlabeled target data and learn domain-invariant representations. Figure 2.13 depicts the process of UDA, where supervised training using semantic annotations occurs in the source domain with a loss function denoted as $\mathcal{L}$. At the same time, unsupervised adaptation is applied on the unlabeled target domain with diverse strategies at the input, feature, or output levels. In this context, commonly employed adaptation strategies encompass discrepancy-based method, pseudo-labeling-based method, adversarial-based method, and auxiliary-modality-based Method.

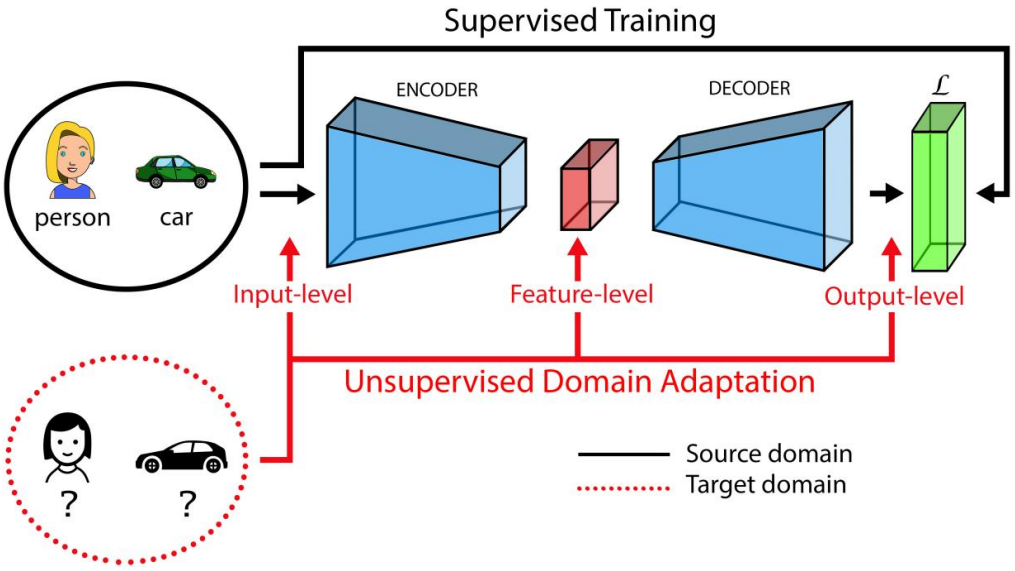


Figure 2.13: Illustration of UDA process.
The image is from [86]

2.4.1 Discrepancy-based Methods

Discrepancy-based methods strive to align the data distributions between two domains. For instance, Maximum Mean Discrepancy (MMD) [87] defines a distance function in the reproducing kernel Hilbert space, which serves to measure the disparity between two distributions. In the context of UDA, this distance can be minimized to align the data distributions of two distinct domains, which can be formulated as:

$$MMD\left(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}\right) = \left\| \frac{1}{\mathcal{N}_{\mathcal{S}}} \sum_{i=1}^{\mathcal{N}_{\mathcal{S}}} \phi\left(\mathcal{X}_{\mathcal{S}}{}^{i}\right) - \frac{1}{\mathcal{N}_{\mathcal{T}}} \sum_{j=1}^{\mathcal{N}_{\mathcal{T}}} \phi\left(\mathcal{X}_{\mathcal{T}}{}^{j}\right) \right\|_{\mathcal{H}}^{2}, \tag{2.4}$$

where $\mathcal{N}_{\mathcal{S}}$ and $\mathcal{N}_{\mathcal{T}}$ are number of samples in the source and target domain, $\phi(.)$ is the feature mapping and $\mathcal{H}$ denotes the Reproducing Kernel Hilbert Space (RKHS). [88] integrates MMD-based domain confusion loss with classification loss, thereby optimizing the classifier while minimizing domain distribution distances, ultimately enabling the model to learn domain-invariant representations. Building upon this, DAN [89] leverages the multi-kernel MMD method to align feature distributions across different modalities within each fully-connected layer of the model, thereby further diminishing domain differences. Alternatively, [90] defines the CORAL loss as the distance between the second-order statistics of the source and target features, which in conjunction with classification loss, to learn features that work well on the target domain. In this context, [88] integrates MMD-based domain confusion loss with classification loss, thereby optimizing the classifier while minimizing domain distribution distances, ultimately enabling the model to learn domain-invariant representations. Building upon this, DAN [89] leverages the multi-kernel MMD method to align feature distributions across different modalities within each fully-connected layer of the model, thereby further diminishing domain differences. Alternatively, [90] defines the CORAL loss as the dis-

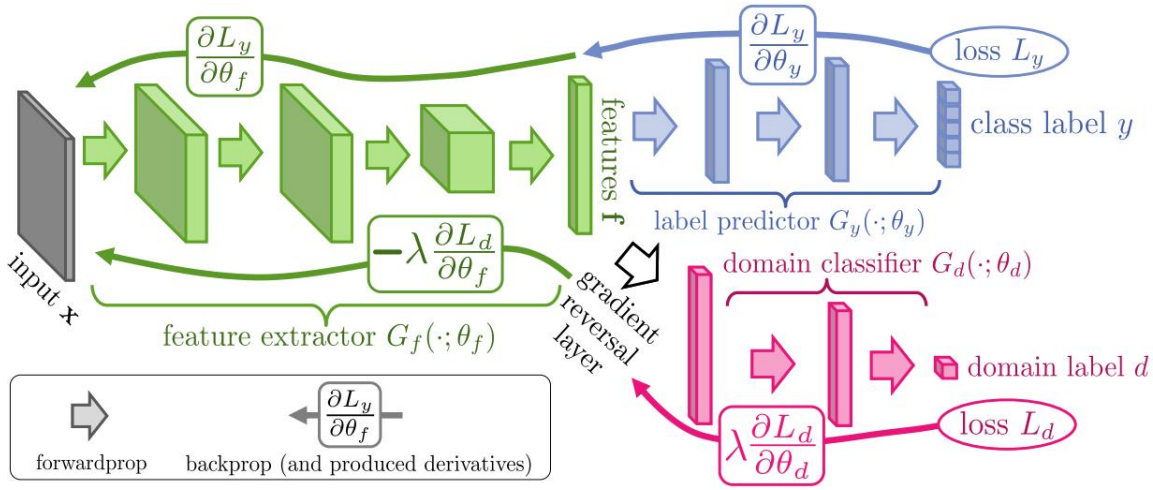


Figure 2.14: Illustration of adversarial training scheme.
The image is from [95]

tance between the second-order statistics of the source and target features, which in conjunction with classification loss, to learn features that work well on the target domain.

2.4.2 Pseudo-labeling-based Methods

Pseudo-labeling-based methods generate pseudo labels for the target domain by predicting the probability distribution of input data on related tasks. In this context, models trained on source domain are often considered pseudo-label generators. For instance, [91] concurrently trains three classifiers, two of which generate pseudo labels on the target domain, and the third leverages these pseudo labels to conduct supervised training. Then, [92] further proposes a progressive feature alignment network, which applies an easy-to-hard transfer strategy and adaptive prototype alignment strategy to select reliable pseudo labels. Furthermore, [93] introduces a new contrastive domain discrepancy objective to minimize the domain discrepancy within the same class and maximize the domain discrepancy between different classes. Additionally, [94] proposes a domain-specific batch normalization layer that assigns different batch normalization parameters to different domains, thus capturing domain-specific information and transforming it into domain-invariant representations and producing more accurate pseudo labels.

2.4.3 Adversarial-based Methods

Adversarial-based methods, inspired by generative adversarial networks (GANs), mitigate differences between domains by integrating an additional adversarial objective and introducing an auxiliary domain discriminator to differentiate between source and target domains. More specifically, During training, a domain discriminator is first optimized using a standard classification loss to make the source and target domains distinguishable. Thus, given a

discriminator D the loss function can be formulated as follows:

$$\min_{\theta_D} \mathcal{L}_{cls} = \mathbb{E}_{\mathcal{X}_\mathcal{S}}[logD(\mathcal{X}_\mathcal{S}))] + \mathbb{E}_{\mathcal{X}_\mathcal{T}}[log(1 - D(\mathcal{X}_\mathcal{T})))]. \tag{2.5}$$

where θ_D is the parameters of discriminator. Subsequently, the parameters of the model, θ_E, is updated following the adversarial optimization process to fool the discriminator, which can be summarized as a min-max criterion:

$$\min_{\theta_E} \max_{\theta_D} \mathcal{L}_{adv}(\mathcal{X}_\mathcal{S}, \mathcal{X}_\mathcal{T}, D) = \mathbb{E}_{\mathcal{X}_\mathcal{T}}[log(D(\mathcal{X}_\mathcal{T}))]. \tag{2.6}$$

In early explorations, [95] integrated a gradient reversal layer that inverted the gradient generated by domain classification loss during backpropagation, aiming to force feature distributions across different domains indistinguishable, as shown in Figure 2.14. Following this, [96] incorporated adversarial training strategies with Maximum Mean Discrepancy (MMD) and proposed a joint maximum mean discrepancy criterion to align the joint distributions of activations across multiple task-specific layers. Then, [97] proposed adversarial discriminative domain adaptation, a simple way that explicitly uses domain adversarial loss to project data from diverse domains into a shared feature space. Moreover, to eliminate low-level disparities between domains, [98] applied pixel-level adaptations and introduced cycle-consistent adversarial domain adaptation, known as CyCADA, which directly remaps the source training data to the target domain, subsequently utilizing the remapped data for adversarial training. Besides, [99] introduced a hybrid adversarial network (HAN) that integrates the adversarial training process based on domain discrimination with feature distribution alignment strategy with CORAL loss, offering a solution for joint adversarial learning with class information and domain alignment.

2.4.4 Auxiliary-modality-based Methods

Recently, several efforts have been made to bridge the domain gap by capitalizing on the complementary information provided by additional modalities, e.g., depth maps. For example, [19] introduced a unified depth-aware UDA (DADA) framework, as shown in Figure 2.15. In this framework, depth information is incorporated into the semantic segmentation model through an auxiliary depth regression task, which in turn provides additional privileged information during the training process. Furthermore, from the perspective of multi-task complementarity, [20] delved deeper into UDA within the framework of multi-task learning and proposed a cross-task relation layer (CTRL), which encodes the task dependencies between semantic segmentation and depth estimation to improve performance in these tasks. In addition, [18] proposed a correlation-aware domain adaptation (CorDA) approach, which employs a domain-shared multi-modal distillation module to model and leverages the correlation between semantics and depth features, thereby guiding the refinement of pseudo labels in the target domain.

Diverging from the methods above, [100] developed geometrically guided input-output adaptation, known as

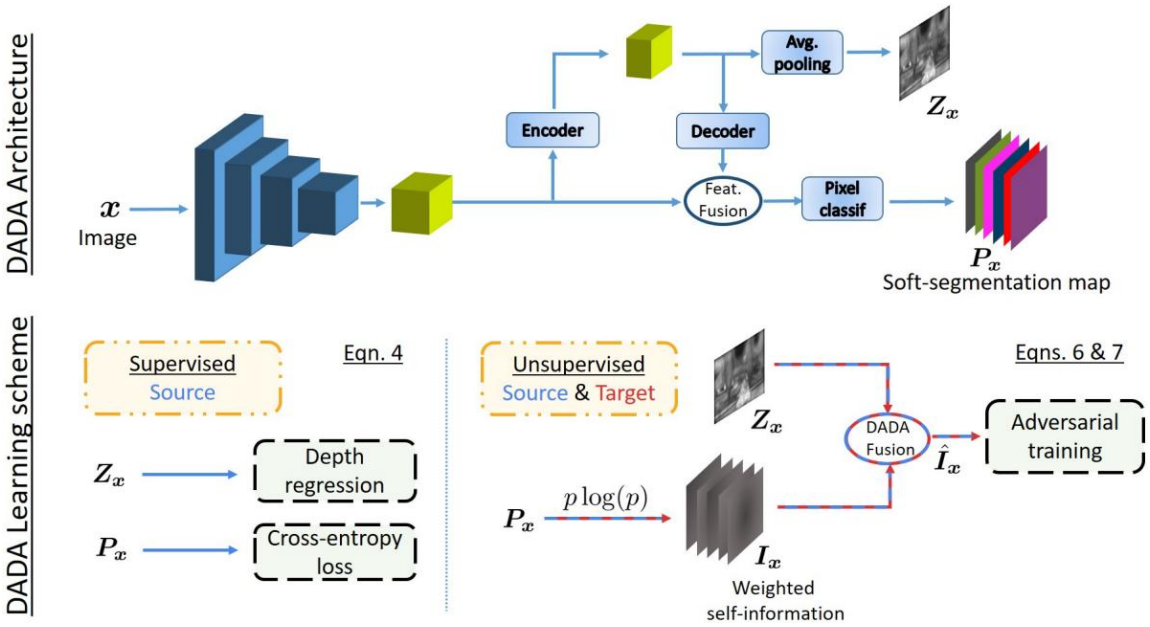


Figure 2.15: Illustration of DADA architecture (top) and DADA learning scheme (bottom). The image is from [19]

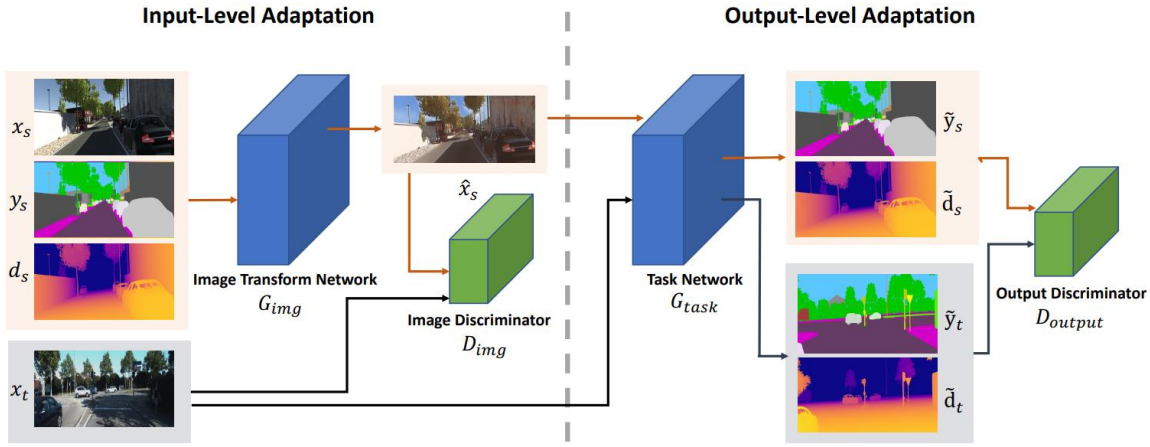


Figure 2.16: Illustration of GIO-Ada architecture. The image is from [100]

GIO-Ada, which integrates adaptations on both input and output levels. As illustrated in Figure 2.16 this approach leverages color, semantic, and geometric cues from the source domain at the input level to establish correlations among RGB images, semantics, and geometry, while also performing depth estimation during domain adaptation.

2.5 Summary

In this chapter, we initially reviewed the progression of DL-based scene perception techniques, with a particular focus on tasks involving semantic segmentation and object detection. Subsequently, we reflected on the strategies related to multi-modal fusion, including early fusion, late fusion, and intermediate fusion. This discussion mainly encompassed methods for the fusion of RGB images with depth maps, as well as the fusion of RGB images with thermal images. Finally, we delved into UDA-related works and reviewed the discrepancy-based, pseudo-labeling-based, adversarial-based, and auxiliary-modality-based Methods. We observed that multi-modal approaches could provide complementary, scene-specific information, thereby bolstering the stability and precision of environmental perception. Furthermore, multi-modal related techniques can be broadly employed in a diverse set of tasks to assist in developing more adaptive and comprehensive models.

Chapter 3

A General Two-Branch Decoder Architecture for Semantic Segmentation

Semantic segmentation is a fundamental task in computer vision that allows a comprehensive assessment of a model's ability to parse a scene. We observed that most multi-modal semantic segmentation approaches are extensions of single-modality-based counterparts. Therefore, to obtain a more thorough understanding of semantic segmentation models, this chapter embarks from a general encoder-decoder architecture and places particular emphasis on exploring single-modality-based, i.e., RGB image, semantic segmentation methods. Subsequently, we attempt to enhance the performance of existing methods by optimizing both the training strategies and the architecture of the models. Specifically, the research conducted in this chapter covers two branches. The first branch seeks to explore the potential of enhancing model segmentation accuracy by leveraging the additional supervisory cues offered by semantic boundaries. The second branch delves into the possibility of enhancing performance by optimizing the structure of the decoder.

3.1 Abstract

Recently, many methods with complex structures were proposed to address image parsing tasks such as image segmentation. These well-designed structures are hardly to be used flexibly and require a heavy footprint. In this chapter we focuse on a popular semantic segmentation framework known as encoder-decoder, and points out a phenomenon that existing decoders do not fully integrate the information extracted by the encoder. To alleviate this issue, we propose a more general two-branch paradigm, composed of a main branch and an auxiliary branch, without increasing the number of parameters, and a boundary enhanced loss computation strategy to make two-branch decoders learn complementary information adaptively instead of explicitly indicating the specific learning

element. In addition, one branch learn pixels that are difficult to resolve in another branch making a competition between them, which promotes the model to learn more efficiently. We evaluate our approach on two challenging image segmentation datasets and show its superior performance in different baseline models. We also perform an ablation study to tease apart the effects of different settings. Finally, we show our two-branch paradigm can achieve satisfactory results when we remove the auxiliary branch in the inference stage, so that it can be applied to low-resource systems.

3.2 Introduction

Semantic segmentation can be formulated as the task of labeling all pixels in an image with semantic classes. Most state-of-the-art semantic segmentation models are based on the encoder-decoder architecture or its variants. Specifically, the encoder extracts information from the original input, and the decoder integrates previously extracted information and recovers semantic information from it. In recent years, researchers commit to exploring different network architecture [39, 9] to learn a more general representation, then deployed to the image segmentation task [48, 101]. However, a general representation extracted by the encoder means that the decoder need to decrease the gap between task free representation and task-dependent information.

In order to improve the parsing ability of the decoder, DeeplabV3+ [48] through pyramid pooling integrate the contextual information at multiple scales. FCN [102] use skip-connection to fuse feature maps of different layers. [103, 45] try to explore the interrelationships between features through attention mechanisms. It is worth noting that some recent works start exploring the two-branch structure in the decoder [104, 49]. They capture meaningful information by carefully designing different branches. Unfortunately, existing two-branch structures were elaborately designed, thus hard to port to other types of decoders, and the degradation of model performance caused by removing a branch is also unacceptable. Or they were just designed for post-processing and are challenging to train end-to-end. On the other hand, with the continuous improvement of the encoder's representation ability, making full use of the information extracted by the encoder is still an open question. Therefore, we have reason to suspect that the existing encoder-decoder-based models do not fully integrate the information extracted by the encoder. We verified this view through experiments.

To alleviate these problems, we propose a more general two-branch paradigm, composed of a main branch and an auxiliary branch for improving the structure of the decoder. At the same time, we design a simple yet efficient branch that can be flexibly integrated into existing encoder-decoder semantic segmentation systems to verify the effectiveness of the proposed two-branch structure. In order to enable two branches to learn complementary information, we customize a loss calculation method to supervise the learning process of each branch. With these ideas, different branches can learn complementary information adaptively instead of explicitly indicating the specific learning elements of different branches. In addition, learning complementary information can make the two branches

compete with each other to a certain extent during the learning process, which can further improve performance. Moreover, compared with the counterpart of the original model, the ameliorated two-branch version reduces or maintains the number of parameters while improving performance.

Our main contributions can be summarized as follows:

- We propose a general two-branch paradigm to enhance the capability of the decoder to parse the information extracted by the encoder without increasing the number of parameters.
- We propose the BECLoss that can supervise two-branch decoders to learn complementary information adaptively instead of explicitly indicating the specific learning elements to each branch.
- We design a simple yet efficient branch that can be flexibly integrated into the existing encoder-decoder framework to form a two-branch structure.

3.3 Related Works

3.3.1 Encoder-decoder and Variants

As a general structural paradigm, encoder-decoder is widely used in the field of image segmentation. Such a structure usually first encode features from the input to a latent feature space, then gradually recover the information in the decoder. U-Net [44] explored the potential relationship between the features of the encoding phase and their counterpart in the decoding phase through multiple skip-connections. SEMEDA [105] first learned to convert the label to an embedding space under the guidance of the boundary information, and then supervised the encoder-decoder structure under the learned subspace. PSPNet [41] and Deeplab family [48, 42] introduced dilated convolution in encoder for increasing the receptive field while maintaining the resolution, then several parallel pyramid pooling were followed to integrate information at different scales. Inspired by [43, 106], attention mechanism and its variants are adopted in encoders or decoders [103, 107] to improve performance. In [45], attention was deployed in the decoding stage for re-calibrating the feature maps with learnable weights. In addition, the application of self-attention [108] in encoder has gradually become popular due to its capability of encoding distant dependencies for better feature extraction. SETR [109] adapted a pure transformer encoder to extract features from an image seen as a sequence of patches then followed a decoder to restore the semantic information.

3.3.2 Multi-branch

Learning different information through multiple parallel data streams has been proved to have more advantages for representation and generalization. Specifically, HRNet [101] repeatedly exchanged the information across different resolutions by a series of parallel feature extraction streams in the encoding process to maintain high-resolution

representations. Based on HRNet, [110] proposed a hierarchical multi-scale attention approach in which each data stream learned a specific image scale so that the model can consider the information of multiple input image scales when predicting. GSCNN [111] designed a two-stream structure, one for context information extraction, another one for boundary-related information extraction. Combined with attention, RAN [112] proposed a three-branch structure that performs the forward and backward attention learning processes simultaneously. Similarly, DANet [104] used a two-branch encoder to learn the semantic relevance in spatial and channel feature spaces respectively. Unlike above works, SegFix [49] proposed a post-processing scheme that predicted boundary and direction maps employing a two-branch decoder supervised by two boundary-related losses.

Encouraged by multi-branch learning, we propose a more general and easy-to-deploy two-branch paradigm, in which a new branch can be easily inserted into the original decoder to form a two-branch decoder and, as a result, improve the discriminating ability. Unlike previous works, we design a general paradigm and enable different branches to learn complementary information adaptively instead of explicitly indicating the specific learning elements of different branches.

3.4 Methodology

In this section, we first systematically describe the two-branch decoder paradigm, then design a simple yet efficient branch that can be applied as a plug-in to existing encoder-decoder frameworks to turn them into our proposed two-branch architecture. Finally, we introduced a new loss calculation method that can be used to supervise branch learning complementary information.

3.4.1 Two-Branch Structure Prototype

In an image segmentation model, existing encoder-decoder architectures can be simply represented in Figure 1.1. Our proposed encoder-decoder based two-branch variant is depicted in Figure 3.1. As shown in Figure 3.1 (a), raw data is first input into the encoder for feature extraction, then encoded features are input to two branches separately, followed by a residual-liked module to integrate information from different branches adaptively. For the fusion of two branch features, we use the output of the penultimate layer of each decoder instead of the last layer to retain more information. Specifically, in the residual path, we first concatenate the output features of two branches, next follow a 1×1 convolution to reduce the channels. Then features are combined with the output of the first branch by an element-wise addition operation. The final output is up-sampled to recover resolution if needed.

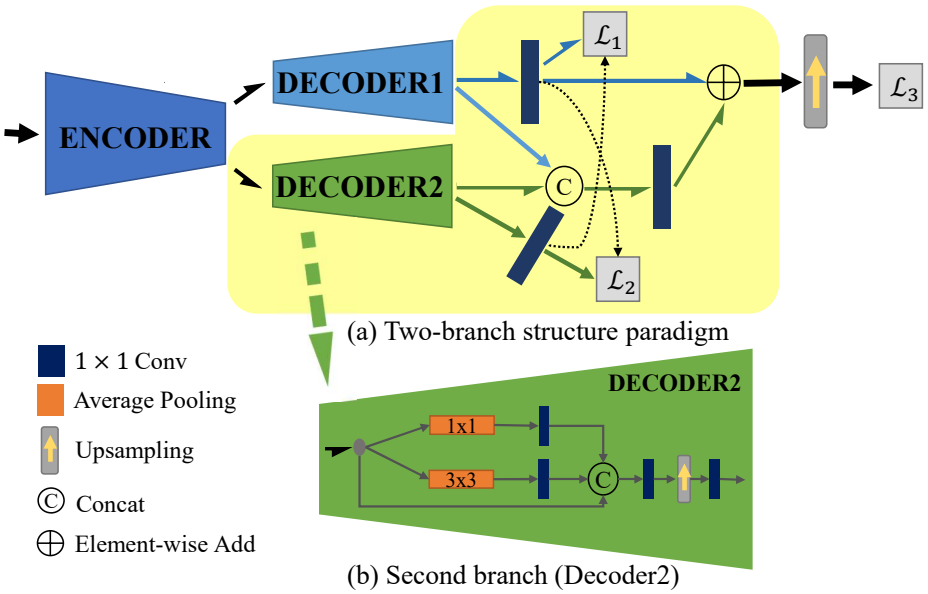


Figure 3.1: Overview of our proposed two-branch architecture. The output of the encoder is divided into two groups, which are represented by two 'half arrows'. Then each group is input to each branch separately and followed by a residual-liked module to fuse the outputs of two branches.

3.4.2 Additional Branch Setting

In this part, we design a simple branch that can be deployed into an encoder-decoder framework to form a two-branch decoder architecture. As shown in Figure 3.1 (b), the branch takes the encoded features as input. Similarly to [41], we utilize a parallel average pooling module, each path consisting of an average pooling operator and a 1×1 convolution operator. We concatenate the output of each path to get a multi-scale feature representation and followed by another 1×1 convolution. Then, we get the output of this branch through an up-sampling operation and a 1×1 convolution operation. Finally, we divide the encoded features into two groups along the channel axis, and each grouped feature is entered into a specific branch.

3.4.3 BECLoss

In supervised learning, loss function plays a crucial role in the optimization of the network. Thus, we further propose a novel loss computation strategy that can efficiently optimize this two-branch structure. Moreover, [105, 111] have proved that introducing boundary information in the loss helps to improve the inherent sensitivity of the network to boundary pixels. Thus, we introduce boundary information in the proposed loss to help the model learn boundary features during the training stage, which is verified in ablation experiments.

We name this well-designed loss BECLoss. Specifically, BECLoss takes three inputs: outputs of the first branch X^1 and the second branch X^2 and ground-truth map GT. We assume batch size as 1, thus the shape of $X^k(k = 1, 2)$ is $C \times H \times W$ and C, H and W indicate the number of predicted classes, high and width of input images, respectively.

First, we get the probability distribution $S^k \in \mathbb{R}^{H \cdot W \times C}$ which can be computed as:

$$S_i^k = \frac{exp(X_i^k)}{\sum_j^C exp(X_i^k[j])}, \tag{3.1}$$

where $i = 0 \dots H \times W - 1$ denotes the index of pixels, $j = 0 \dots C - 1$ denotes the index of channels. Then, we compute the probability map of ground truth label $P^k \in \mathbb{R}^{H \cdot W \times 1}$ as:

$$p_i^k = S_i^k[gt_i] \tag{3.2}$$

where gt_i is the i^{th} pixel in GT. Following, we define a mask M^1 for indicating all the pixels whose probability in P^1 is less than a threshold τ. M^1 indicates the pixels that are difficult to predict in the first branch. With the computed M^1 and P^2, we filter out all pixels in X^2 whose probability is less than a threshold τ:

$$M_i^1 = \begin{cases} 1 & \text{if } P_i^1 < \tau \\ 0 & \text{otherwise} \end{cases}, \tag{3.3}$$

where $i = 0 \dots H \times W - 1$ denotes the index of pixels.

In order to standardize the loss definition, we use $\mathcal{L}^1$ to indicate the boundary enhanced loss computed from X^1, and $\mathcal{L}^2$ to indicate a partial loss that we get from X^2. In $\mathcal{L}^1$ and $\mathcal{L}^2$ we only consider the pixels which are hard to predict in the first branch in order to utilize the additional branch to assist in the prediction of these pixels. In addition, we use a hyperparameter γ to control the influence of boundary information $B \in \mathbb{R}^{W \times H}$ (detailed in 3.4.4) to the loss of the first branch, we get $\mathcal{L}^1 \in \mathbb{R}^{H \cdot W \times 1}$:

$$\mathcal{L}_i^1 = -log(P_i^1) \times (1 + \gamma \cdot B_i) \times M_i^1 \tag{3.4}$$

where $i = 0 \dots H \times W - 1$ denotes the index of pixels. Following, we compute the partial loss $\mathcal{L}^2 \in \mathbb{R}^{H \cdot W \times 1}$:

$$\mathcal{L}_i^2 = -\log(P_i^2) \times M_i^1 \tag{3.5}$$

Finally, the BECLoss can be written as a weighted average sum of $\mathcal{L}^1$ and $\mathcal{L}^2$:

$$\mathcal{L}_{BEC} = \frac{\sum_i (\mathcal{L}_i^1 + \eta \cdot \mathcal{L}_i^2)}{\sum M_i^1} \tag{3.6}$$

where η is a hyperparameter used to control the ratio of $\mathcal{L}^2$ in $\mathcal{L}_{BEC}$.

The two branches can automatically learn complementary information which helps the proposed model to further learn a more appropriate way to combine the outputs of the two branches.

3.4.4 Ground-Truth Boundary

In this part, we explain how we get a ground-truth boundary map from a ground-truth label map. Introducing approximate boundary information in the loss can improve the model's sensitivity to physical boundaries, which improves the prediction accuracy in the boundary area. However, there are always labeled error pixels in the hand-labeled ground truth map, which are especially obvious at the boundary region, as shown in Figure 3.2(a). In order to alleviate this problem, Figure 3.3 illustrates the inner boundary extraction process. Concretely, we first extract the boundary map B^* from the original ground-truth label map by a filter f that sets all pixels that do not have 8 identically-labeled neighbor pixels as 1, and other pixels as 0. Then we thicken the boundary by a 7×7 dilation operator and get boundary map B_t^*. Finally, we get the inner boundary B_{in}^* by applying the same filter f on B_t^* again and followed by another 3×3 dilation operator, as shown in Figure 3.2(b).

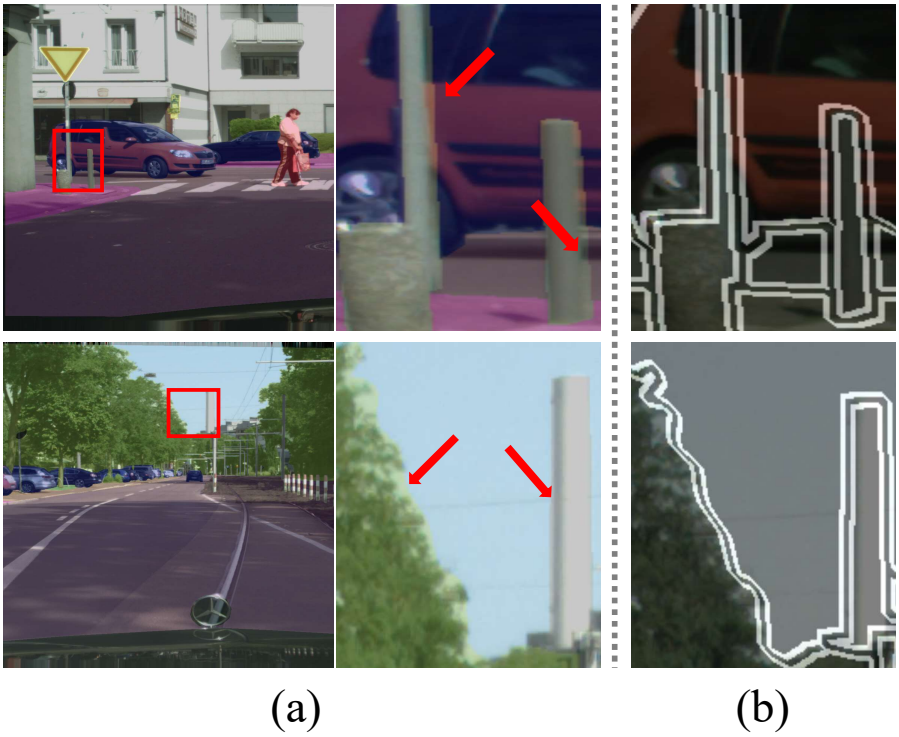


Figure 3.2: (a) Mis-labeled boundary pixels and (b) Extracted inner boundary.

3.4.5 Joint Loss

The proposed BECLoss is designed for optimizing the network with two branches. The purpose is to guide the two branches to learn complementary information. It can naturally be combined with other losses for training the whole network. Therefore, the network is trained to minimize a joint loss function:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \cdot \mathcal{L}_{BEC1} + \beta \cdot \mathcal{L}_{BEC2} \tag{3.7}$$

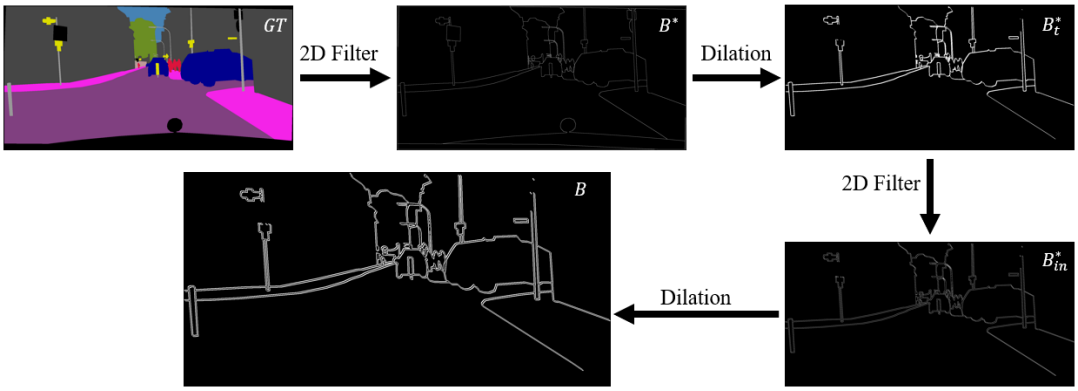


Figure 3.3: Ground-truth inner boundary extraction process.

Specifically, $\mathcal{L}_{CE}$ is cross-entropy loss, $\mathcal{L}_{BEC1}$ and $\mathcal{L}_{BEC2}$ are proposed BECLoss for first and second branch, respectively. α and β are weights parameters of the two BECLoss.

3.5 Experiments

In this section, we conduct experiments on Cityscapes dataset [31] and Freiburg Forest dataset [113]. In the following, we first modify some classic image semantic segmentation algorithms to build their two-branch decoder counterpart, then compare the proposed two-branch architecture with the original network. Finally, we carry out a series of ablation experiments on Freiburg Forest dataset. Our models are trained on one Nvidia Tesla P100 GPU with mixed precision settings.

3.5.1 Datasets

Cityscapes. The Cityscapes dataset is a large-scale database for urban street scene parsing. It contains 5000 finely annotated images captured from 50 cities with 19 semantic object categories, in which 2875 images are used for training, 500 and 1525 images are used for validation and testing separately. All images are provided with a resolution of 2048×1024. We followed [81] and report results on the reduced 11 class label set.

Freiburg Forest. The Freiburg Forest dataset is an unstructured forested environments dataset. It contains 6 segmentation classes, i.e., sky, trail, grass, vegetation, obstacle, and void. The dataset contains 325 images with pixel level hand-annotated ground truth map. We follow [81] and use the same train and test splits provided by the dataset.

3.5.2 Implementation Details

In order to comprehensively test, we deploy proposed two-branch decoder on three classic baseline networks, namely, SegNet [38], DeeplabV3+ [48], and HRNet [101]. Two-branch SegNet is shown in Figure 3.4. We divide

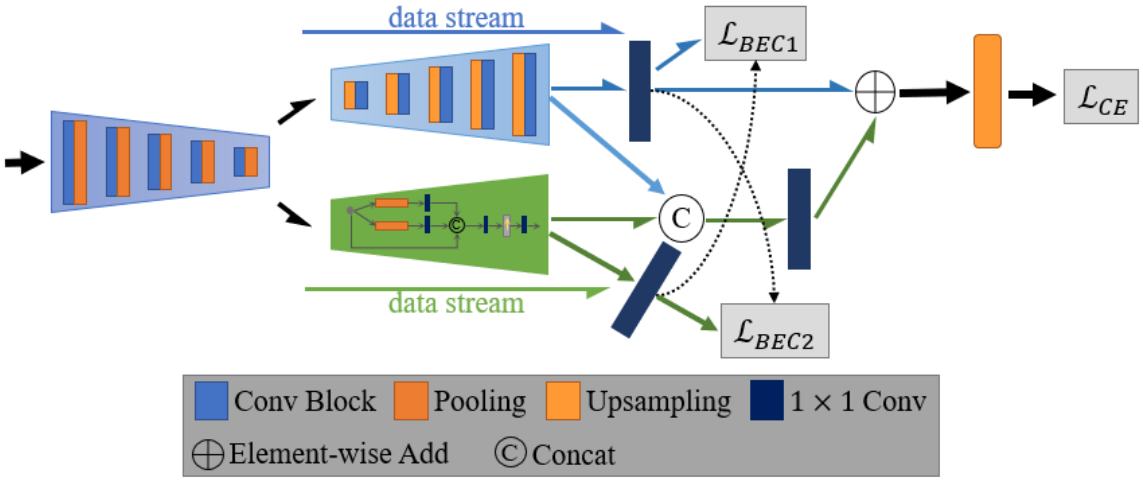


Figure 3.4: Architecture of modified SegNet with two decoders (SegNetT).

the output of the encoder into two groups, one of which is input to the original data stream, and another is input to the additional data stream. In our two-branch implementation, we denote the upper branch in the decoder as the original data stream, the lower branch as the additional data stream. Next, we follow the residual-liked module to fuse the two outputs while deploying the BECLoss and cross-entropy loss during the training. More concretely, we supervise the learning process of the two branches through $\mathcal{L}_{BEC1}$ and $\mathcal{L}_{BEC2}$, and the combination of two outputs are guided by $\mathcal{L}_{CE}$. We follow the same way to implement the counterpart of DeeplabV3+ and HRNet. Note that we only take the backbone in the original model as an encoder, and the rest as the decoder. In practice, we use Resnet50, Vgg16 and HRNet-W18 as backbones.

We initialize encoder with the weights pre-trained on ImageNet, this is totally the same as its original implementations [38, 48, 101]. We employ a cyclical exponent learning rate policy [114] where the min_lr and max_lr are set to $1e-5$ and $1e-2$, and cycle_length and step_size are set to 40 and 5 epochs respectively. Momentum and weight decay coefficients are set to 0.9 and 0.0005. If not specified, all models are trained with a mini batch size of 8. Furthermore, we configure the hyperparameter γ and η in BECLoss as 10.0 and 0.3. The scale α and β in Equation 3.7 are set to 2.0. For Cityscapes dataset, we set input image size to 384×768, thus random cropping (cropsize 384×768) is applied during training, and during testing, we use the original resolution of 1024×2048. For Freiburg Forest dataset, we resize the image to 384×768 during training and testing. All training images are augmented by random left-right flipping. We set 160 and 120 training epochs to Cityscapes datasets and Freiburg Forest dataset. In addition, as we compare the original models with their two-branch encoder counterpart, so we perform the same settings for each comparison pair to ensure fairness.

Table 3.1: Improvements with two-branch decoder on Cityscapes val set with 11 semantic class labels.

Methods	BaseNet	Mean IoU (%)	Parms. (M)
SegNet	Vgg16	75.82	29.4
SegNetT (ours)		80.64 (+**4.82**)	18.6
DeepLabv3+	Res50	80.31	26.6
DeepLabv3+T (ours)		82.45 (+**2.14**)	27.5
HRNet-W18	Hrnet-W18	82.34	9.6
HRNet-W18T (ours)		83.9 (+**1.56**)	9.6

Table 3.2: Comparison in terms of IoU vs different baselines on the cityscapes val set with 11 semantic class labels.

Methods	sky	building	road	sidewalk	fence	vegetation	pole	vehicle	traffic sign	person	bicycle
SegNet	91.83	88.47	95.52	72.76	40.02	91.22	52.95	89.45	65.57	77.2	68.98
SegNetT (ours)	**93.35 (+1.52)**	**90.89 (+2.42)**	**96.65 (+1.13)**	**77.28 (+4.52)**	**49.72 (+9.7)**	**92.37 (+1.15)**	**61.54 (+8.59)**	**92.86 (+3.41)**	**75.64 (+10.07)**	**81.61 (+4.41)**	**75.06 (+6.08)**
DeepLabv3+	93.99	90.9	97.29	80.47	54.73	91.92	56.56	93.05	71.78	78.9	73.79
DeepLabv3+T (our	93.95	**91.99 (+1.09)**	**97.62 (+0.33)**	**82.31 (+1.84)**	**54.85 (+0.12)**	**92.55 (+0.63)**	**62.69 (+6.13)**	**94.17 (+1.12)**	**77.87 (+6.09)**	**82.4 (+3.5)**	**76.59 (+2.8)**
HRNet	94.31	92.06	97.67	82.31	54.94	92.59	63.31	94.34	76.37	82.27	75.4
HRNet-T (ours)	**94.83 (+0.52)**	**92.68 (+0.62)**	**97.98 (+0.31)**	**84.3 (+1.99)**	**56.28 (+1.34)**	**93.06 (+0.47)**	**67.17 (+3.86)**	**94.83 (+0.49)**	**79.98 (+3.61)**	**84.35 (+2.08)**	**77.09 (+1.69)**

3.6 Results

In this section, we provide an extensive evaluation of each component of our framework on two challenging outdoor datasets, namely Cityscapes dataset and Freiburg Forest dataset. We use the widely used intersection over union (IoU) to evaluate the performance of our approach.

3.6.1 Results on Cityscapes Dataset

Table 3.1 summarizes the results of our two-branch decoder with different baselines. We can see that our approach significantly improves the mean IoU. Specifically, our approach improves the mean IoU of original encoder-decoder frameworks, namely SegNet, Deeplabv3+, and HRNet, by 4.81, 2.14, and 1.56, respectively. In particular, our two-branch implementation of SegNet (SegNetT) dramatically reduces the number of parameters while significantly improving the performance. DeepLabv3+T and HRNet only slightly increase the parameters (0.9M) or keep the number of parameters while improving the model's performance. Our results also reflect that the original decoder does not fully use the information extracted by the encoder. In addition, table 3.2 illustrates the category-wise comparison between various baselines and their two-branch variants. We surprisingly find that our method has a significant improvement in the prediction accuracy of small-scale targets, like "pole", "traffic sign" and "person". Several segmentation results are shown in Figure 3.5, we can see that our two-branch variants perform better on those small-size-object classes in the images than the baseline models.

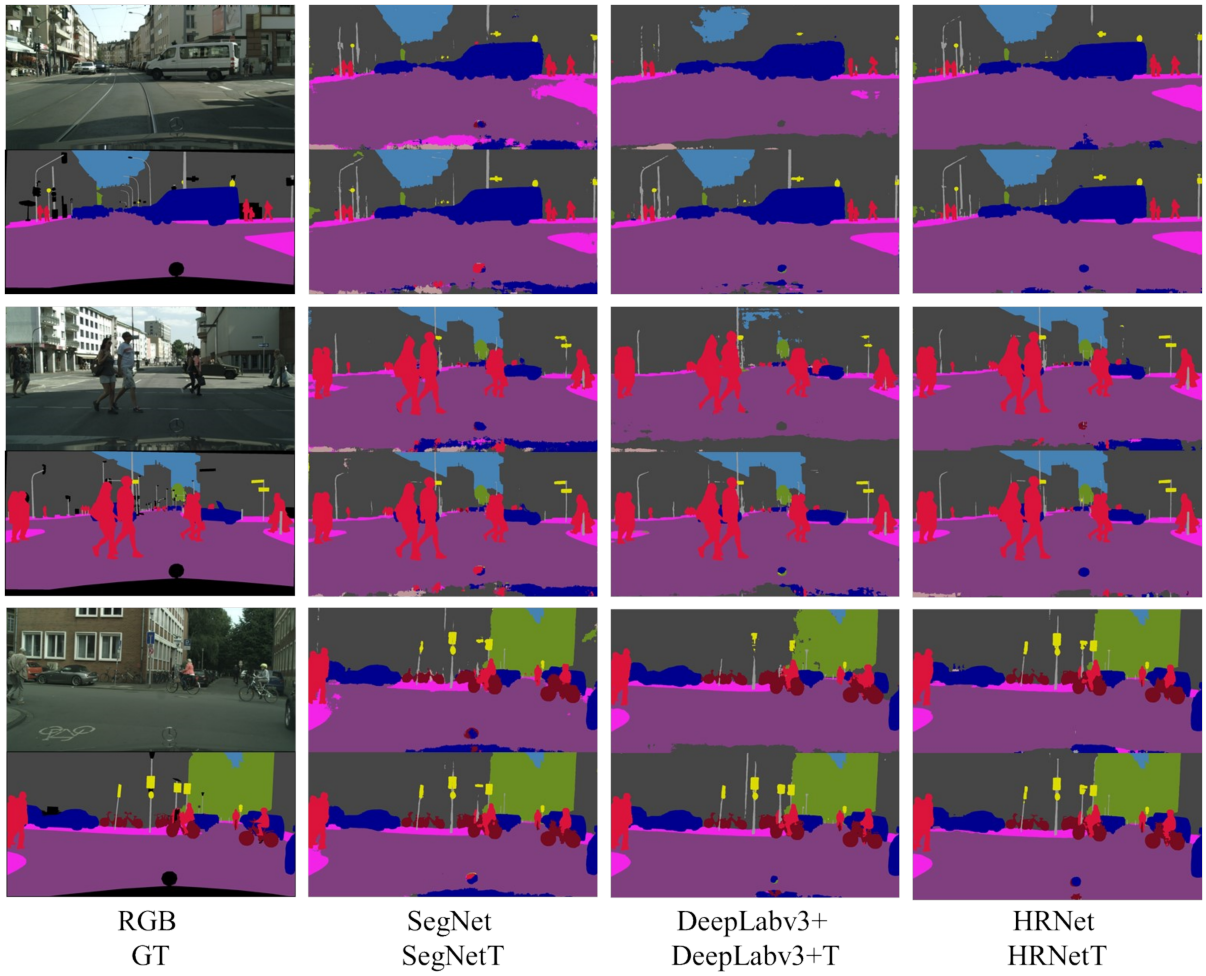


Figure 3.5: Qualitative results on the Cityscapes val set with 11 semantic class labels.

3.6.2 Results On Freiburg Forest Dataset

We carry out experiments on the Freiburg Forest dataset to further evaluate the effectiveness of our method. Quantitative results of Freiburg Forest are shown in Table 3.3. The baselines (SegNet, DeepLabv3+, HRNet) yield mean IoU 69.99%, 77.48%, and 78.29%. Our two-branch counterpart boosts the performance to 81.79%, 82.73%, and 83%. We can see that our methods outperform their baselines with notable advantage, especially for the class of "obstacle", which is hardest to segment because of its severe class imbalance. Several examples are shown in Figure. 3.6.

3.6.3 Ablation Study

3.6.4 BECLoss and Boundary

All two-branch variants are implemented by replacing the decoder of the original network with our proposed two-branch decoder, and through our well-designed BECLoss to explicitly supervise the learning process of the model, the two branches can learn complementary information. In addition, we introduce boundary information into BECLoss

Table 3.3: Improvements with two-branch decoder on Freiburg Forest val set.

Methods	BaseNet	Trail	Grass	Veg.	Sky	Obst.	Mean IoU (%)	Parms. (M)
SegNet	Vgg16	84.15	85.55	88.97	91.28	0	69.99	29.4
SegNetT (ours)		**88.55 (+4.4)**	**88.96 (+3.41)**	**0.91 (+1.94)**	**2.63 (+1.35)**	**47.93 (+47.93)**	**81.79 (+11.8)**	18.6
DeepLabv3+	Res50	83.03	86.11	89.96	92.16	36.1	77.48	26.6
DeepLabv3+T (ours)		**88.02 (+4.99)**	**88.93 (+2.82)**	**91.02 (1.06)**	**2.83 (+0.67)**	**52.87 (+16.77)**	**82.73 (+5.25)**	27.5
HRNet	Hrnet-W18	84.79	86.49	89.79	91.96	38.44	78.29	9.6
HRNet-T (ours)		**88.74 (+3.95)**	**89.35 (+2.86)**	**91.14 (+1.35)**	**92.6 (+0.64)**	**53.17 (+14.73)**	**83 (+4.71)**	9.6

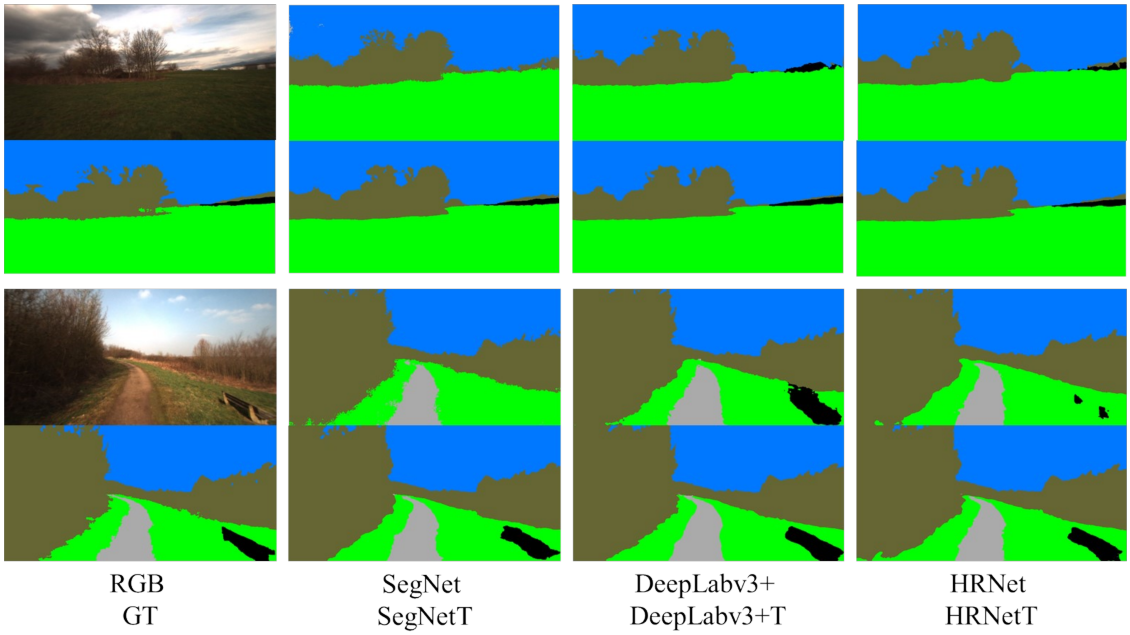


Figure 3.6: Qualitative results on the Freiburg Forest test set.

to improve the inherent sensitivity of our models to boundary pixels. To verify the validity of our method, we conduct a group of ablations to analyze the influence of various factors within our method. We report the results over the segmentation baseline SegNet on Cityscapes and Freiburg Forest dataset in Table 3.4.

As shown in Table 3.4, two-branch decoder improves the performance remarkably. Compared with the baseline SegNet, employing two-branch decoder yields a result of 78.54% mean IoU on Cityscapes dataset and 78.9% mean IoU on Freiburg Forest dataset, which brings 2.72% and 8.91% improvement. In addition, when we gradually replaced the cross-entropy loss CELoss of loss1 and loss2 with the BECLoss we designed, the performance further improved to 79.5% and 81.43%. Furthermore, we notice that when we use only one BECLoss, the result very slightly exceeds the result of using two BECLoss, as shown in the third row and the fifth row, the result from 79.54% goes to 79.5% on Cityscapes dataset. After introducing boundary information to BECLoss, performance further increased to 80.64%. Results show that our proposed two-branch decoder and boundary enhanced BECLoss bring great benefit to scene parsing.

Table 3.4: Ablation study on Cityscapes val set and Freiburg Forest test set. $Loss1$-$Loss3$ represent deployed loss in Figure 3.1, B indicates BECLoss enhanced by boundary information.

Methods	Loss1	Loss2	Loss3	B	Mean IoU (%)	
					Cityscapes	Freiburg
SegNet	\	\	CE	\	75.82	69.99
SegNetT	CE	CE	CE	\	78.54 (+2.72)	78.9 (+8.91)
SegNetT	BEC	CE	CE	N	79.54 (+3.72)	80.48 (+10.49)
SegNetT	CE	BEC	CE	N	79.07 (+3.25)	79.9 (+9.91)
SegNetT	BEC	BEC	CE	N	79.5 (+3.68)	81.43 (+11.44)
SegNetT	BEC	BEC	CE	Y	80.64 (**+4.82**)	81.79 (**+11.8**)

Table 3.5: Single branch test on Cityscapes val set with 11 semantic class labels. '$Enc.$' represent encoder, '$Dec.$' represent decoder. 'O' indicates the decoder deployed in the original model. 'D' the decoder in Figure 3.1(b), 'T' indicates our two-branch decoder. 'O^*', 'D^*' and '$O\&D$' mean the result from upper branch, lower branch and final branch separately.

Methods	Enc.	Dec.	Mean IoU (%)			Parms. (M)		
SegNet	Vgg16	O	75.82			29.4		
ED		D	65.34			15.3		
SegNetT (ours)		T	O^*	D^*	$O\&D$	O^*	D^*	$O\&D$
			80.49	67.34	80.64	18.4	14.9	18.6
DeepLabv3+	Res50	O	80.31			26.6		
ED		D	76.67			32.2		
DeepLabv3+ (ours		T	O^*	D^*	$O\&D$	O^*	D^*	$O\&D$
			82.35	77.3	82.61	25.3	25.7	27.5
HRNet	Res50	O	82.34			9.6		
ED		D	81.25			9.7		
HRNet-T (ours)		T	O^*	D^*	$O\&D$	O^*	D^*	$O\&D$
			83.87	76.83	83.9	9.6	9.6	9.6

3.6.5 Single Branch

As mentioned in section1, the proposed two branches can compete during the training process, which prioritizes each branch to learn complementary knowledge that can boost the parsing ability and improve learning efficiency. Thanks to this property, the results are still far better than the original encoder-decoder structure even if we remove a branch during the inference process. Moreover, the number of parameters is less than the original one, which alleviates the challenging to deploy complex models into practical applications in many real scenarios due to computer resources and run-time limitations. As shown in Table 3.5, we use an extremely simple branch, illustrated in Figure 3.1(b), retraining on the Cityscapes dataset, and we named the trained model 'ED'. Moreover, we test the output results of each branch separately on the trained two-branch decoder model. Specifically, we take SegNet as an example. In the inference process, we only keep the upper branch of the model in Figure 3.4, and the output result obtained corresponds to 'O^*'. 'D^*' corresponds to the result of only keep the lower branch. '$O\&D$' goes to the result of original two-branch model. The results of the upper branch in our trained two-branch model are 80.49%, 82.35%, and 83.87%, which significantly exceeds the counterparts of original encoder-decoder models (75.82%, 80.31%, and 82.34%). The results of lower branch trained in two-branch model are also better than correspond one-branch trained model. At the same time, the number of parameters used dropped remarkably. In addition, we find that

the residual-like module can effectively combine the outputs of the two branches to further improve the final result to 80.64%, 82.61%, and 83.9%, as shown in '$O\&D$' columns, which means that the final results are not adversely affected. The results once again show that our method can make each branch learns complementary information.

3.7 Summary

In that chapter, we delve into the RGB-based semantic segmentation models. One of our focuses is enhancing the model's sensitivity towards boundaries. To achieve this, we utilize semantic boundary cues within the loss function, which provides an additional supervisory signal to deliver more precise and effective segmentation. In addition, we also focus on the structure of the decoder and design a generic two-branch decoder that could be flexibly applied to existing encoder-decoder-based models and obtain consistent performance gains.

Specifically, we present a general two-branch decoder paradigm composed of a main branch and an auxiliary branch for scene segmentation. This decoder paradigm can be directly applied in an encoder-decoder framework to efficiently refine and integrate the information extracted by the encoder. With this two-branch decoder, we further propose a boundary enhanced complementary loss named BECLoss to guide two branches to learn complementary information. Moreover, we design a simple yet efficient branch deployed as the auxiliary branch in our two-branch decoder. The comparative experiments show that the proposed two-branch decoder paradigm and BECLoss can significantly improve the performance of the original encoder-decoder model consistently on challenging outdoor datasets. In addition, although we add a branch to the decoder, it does not significantly increase the number of parameters, and the added branch can be removed in the inference process while still getting performance far beyond the original counterpart.

Chapter 4

A Hybrid RGB-D Cross Fusion Network for semantic segmentation

In the previous chapter, we delved into the working paradigm of single-modality semantic segmentation models, i.e., encoder-decoder-based models. Furthermore, in Chapter 2, we provided an overview of methods based on multi-modal fusion, introducing the commonly employed fusion strategies: early fusion, late fusion, and intermediate fusion. Empirical evidence has demonstrated that fusion systems can enhance feature representations of the same scene input data by leveraging the information of different modalities. However, the challenge of how to utilize auxiliary modalities to augment the scene understanding capabilities of DL models remains a complex issue. On the other hand, gradient backpropagation-based DL models are inherently flexible in terms of structural design, which allows us to explore the impact of various fusion strategies on multi-modal inputs.

In this context, this chapter delves into the intricacies of different fusion methods and focuses extensively on how to leverage various intermediate fusion strategies to optimize the feature representations of multi-modal data and subsequently enhance the accuracy and efficiency of semantic segmentation models.

4.1 Abstract

Multi-modal scene parsing is a prevalent topic in robotics and autonomous driving since the knowledge of different modalities can complement each other. Recently, the success of self-attention-based methods has demonstrated the effectiveness of capturing long-range dependencies. However, the tremendous cost dramatically limits the application of this idea in multi-modal fusion. To alleviate this problem, this chapter designs a multi-modal cross-fusion block (AC) and its elegant variant (EAC) based on an additive attention mechanism to capture global awareness among different modalities efficiently. Moreover, a simple yet efficient transformer-based trans-context block (TC)

is also presented to connect the contextual information. Based on the above components, we propose light HCFNet, which can explore long-range dependencies of multi-modal information while keeping local details. Finally, we conduct comprehensive experiments and analyses on both indoor (NYUv2-13, -40) and outdoor (Cityscapes-11) datasets. Experiment results show that the proposed HCFNet achieved 66.9% and 51.5% mIoU on NYUv2-13 and -40 classes settings, which outperform current start-of-the-art multi-model methods. Our model also shows a competitive mIoU of 80.6% on the Cityscapes-11 dataset.

4.2 Introduction

As a fundamental task, semantic segmentation has received a broad range of attention in the computer vision community and industry. Depth information as an auxiliary provides shape and geometry cues of the surroundings that complement the RGB data, thus introducing depth information to improve the model's performance has become a trend in robotics and autonomous driving. To this end, a series of networks with RGB-D as input appeared. These methods directly concatenate RGB and Depth images [71, 115, 70] or treat them in two branches [116, 117, 118, 82].

Recently, self-attention-based transformer architecture has attracted attention in the computer vision community due to its flexibility in long-range modeling dependencies and its remarkable success in natural language processing (NLP). Therefore, well-designed transformer block (TB) or their variants are introduced to replace the region-wise convolution structure [119, 120], and their results have shown that a global view brought by self-attention helps draw a better performance. However, the tremendous computation severely restricts its application in computer vision, especially on some resource-limited and low-latency systems. To alleviate this shortcoming, some works were designed to process images at a low resolution [121, 119] or using a sliding window [120, 122]; others borrowed the idea from CNN-like architecture and introduce pyramid structure [123, 124]. Although various solutions for building long-range dependencies of individual RGB image reasoning emerged, the exploration of fusing RGB and Depth data for scene parsing is very limited. Existing multi-modal methods mainly deploy TB in a hybrid structure, i.e., mixing CNN and transformer, to ingest the advancement of both convolution and TB. The mainstream combinations of CNN and TB operations are (1) cascade: convolution operations are used to process high-resolution data and then followed by TBs to process low-resolution data. (2) parallel: CNNs are used as the backbone network for feature extraction, while TBs are independent modules to address the data fusion and exchange between different modalities. For example, [125] introduces the TB into the last two layers of the encoder in U-Net [44], reducing the calculation amount. [126] treat TB as an independent part and stack multiple self-attention modules to incorporate the global attention of the 3D scene. Compared with parallel structure, cascade structure fails to capture a larger context at the shallow level, and all existing methods rudely use the TB, so they inevitably bear the burden of the TB, i.e., the tremendous calculation.

We argue that the global attention founded by TB is the crucial reason for the success of Transformer. Recently,

[127] proposed an efficient Transformer variant based on additive attention to achieve global attention modeling in linear complexity. Inspired by this, we propose a well-designed additive-attention-based cross-fusion block (AC) to incorporate depth information into RGB and form long-range dependencies between depth and RGB features. Besides, we present EAC block, an efficient variant of AC, which efficiently builds global contexts while maintaining fine-grained shape details. On the other hand, we offer a simple yet efficient trans-context module (TC) to enrich contextual information and capture a global context from fused features. Based on the above modules, we design a hybrid cross fusion network (HCFNet), as shown in Figure 4.1. With all the ideas, our method benefits from building global awareness while significantly reducing computational consumption. The formed global awareness crosses RGB and depth, bringing integrated information from different modalities. We report the experimental results on two commonly used datasets, namely NYUv2(-13, -40) [128] and Cityscapes-11 [31], to verify the effectiveness of the proposed method in both indoor and outdoor scenarios.

The main contributions of this chapter are summarized as follows:

- We propose an efficient hybrid RGB-D data fusion network called HCFNet for semantic segmentation.
- We propose a light data fusion block named additive attention cross fusion block (AC), and its variant (EAC), to form long-range dependencies cross depth and RGB features. Moreover, we offer a simple yet efficient trans-context module (TC) based on TB to build a global view of fused features.
- We experimentally validate the proposed HCFNet on indoor and outdoor datasets, including NYUv2(-13, -40) and Cityscapes-11. Results show that our method achieves 66.9% mIoU on NYUv2-13, 51.5% mIoU on NYUv2-40, and 80.6% mIoU on Cityscapes-11 dataset, which is quite competitive compared with state-of-the-art RGB-D fusion methods.

4.3 Related Works

4.3.1 Global Attention and Transformer

Transformer was firstly proposed by [108] for NLP tasks. The core component is built upon multi-head self-attention, which can model the long-range dependencies within a sequence. A similar idea was introduced by [129] in computer vision to design a non-local block to build the global relationship between pixels. [130] proposed a full attention block based on a non-local block that computes global attention along both channel and spatial dimensions. As a pioneer of visual Transformer, [121] used pure Transformer structure to classify images and achieved promising results. [119] extended the work in [121] to semantic segmentation. Then, [120, 122] calculate self-attention in sub-windows to alleviate the resolution disaster, and apply Transformer structure to dense segmentation. [123, 124, 131] employed a pyramid or hierarchical Transformer structure to improve the computational efficiency of the model

for segmentation. Moreover, [132] proposed a 'transposed' self-attention that computes global attention across feature channels so that the computational complexity is linear. Recently, [127] offered a variant of Transformer, in which additive attention replaces self-attention to establish global awareness. Compared with other global attention mechanisms, the calculation of additive attention is more efficient, so this article establishes a more general cross-modal fusion attention mechanism based on additive attention.

4.3.2 RGB-D Semantic Segmentation

With a more affordable depth sensor, semantic segmentation leveraged by the complementary geometric information of depth has drawn attention. However, the large noise in depth and the asymmetry between RGB and depth data make it challenging to integrate RGB and depth features effectively. In general, existing semantic segmentation structures include two stages: encoding and decoding. Concretely, input data are first encoded to form contextual feature embeddings then decoded to recover semantic information [14]. Some work [133, 134, 70] redesigned the convolution operation based on the characteristics of RGD-D data. [135] presented depth-aware operations to leverage depth similarity between pixels. [70] proposed a shape-aware convolutional layer. This convolutional layer is composed of two independent learnable components in the learning phase, and all the learnable parameters in the inference phase can be re-weight into a standard convolution operation. [133] introduced malleable 2.5D convolution to learn the receptive field along the depth axis. In contrast, most approaches are proposed to feed RGB and depth to two parallel branches [78, 117, 82, 116, 136]. For example, [137] employed two separate encoder-decoders to process RGB and depth, respectively, during which the manually designed gated fusion layer is used to fuse information from different streams. [138] used skip-connection to transmit the encoded multi-modal information to the decoder. [118, 117, 82] fuse the features at different stages of the encoding process. Recently, [136] applied a shallow encoder and factorized convolutions to create a lightweight model for real-time operations.

Unlike the above methods, we design a hybrid cross fusion network that takes advantage of long-range dependencies in the Transformer while maintaining the model's efficiency.

4.4 Methodology

4.4.1 Overview

An overview of our hybrid cross fusion network (HCFNet) is presented in Figure 4.1. The structure is derived from a general and classical multi-modal semantic segmentation paradigm, i.e., two encoders for extracting features from RGB and Depth and one decoder for reconstructing features from embeddings. Similar to [117, 136], we use independent modules to achieve data fusion of different modalities and pass features to the decoder via skip-connections. The decoder is divided into multiple stages. In each stage, feature maps are first treated by a series of residual blocks

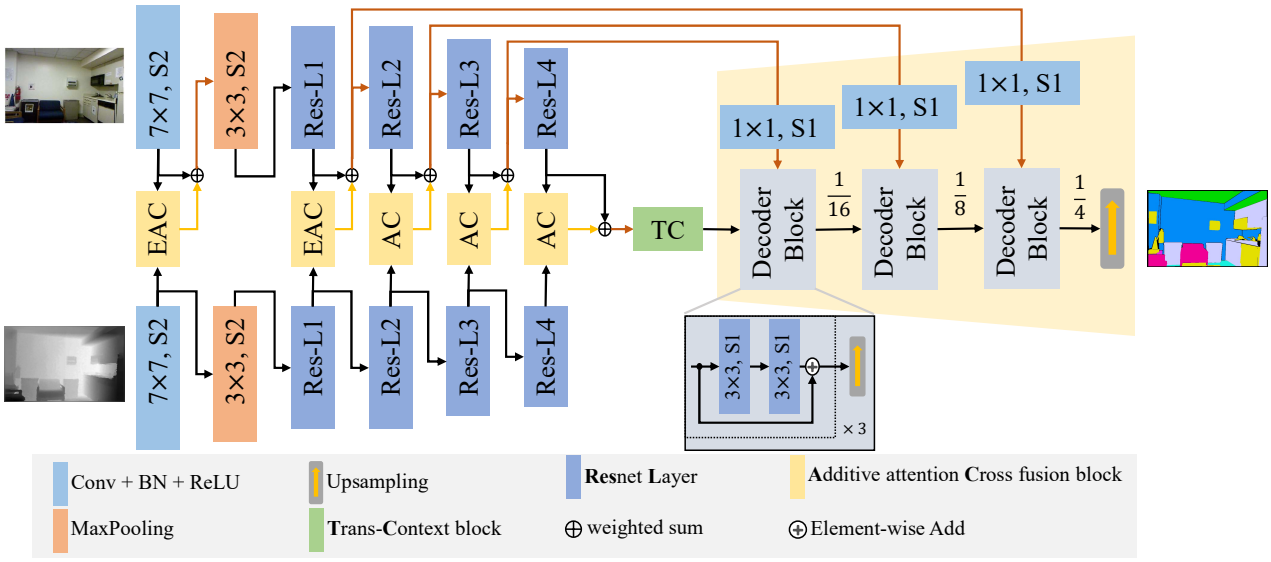


Figure 4.1: Structure of HCFNet. This network takes two inputs, i.e., RGB and Depth. $7 \times 7, S2$ means convolution with kernel size 7 and stride 2, and BN denotes batch normalization.

[9] and then upsampled by a factor of 2. The final output of the decoder is upsampled by the factor of 4 to recover the original resolution. Our network uses shallow encoders (i.e., ResNet-34 [9]) as the backbone for feature extraction of both RGB and Depth streams to reduce the footprint at runtime. In addition, we introduce additive attention cross fusion blocks (AC) and EAC to fuse valuable information efficiently during encoding and trans-context block (TC) to enrich contextual features at the end of the encoder.

4.4.2 Additive Attention

Additive attention was first introduced in [127], which brings an effective global attention mechanism to recalibrate the features within a sequence. A basic form of additive attention is depicted in Figure 4.2. We first summarize each token (T_i, $i \in [1 \dots N]$) into an attention scores by a linear transformation and a scale factor of $\sqrt{d}$, where d is the number of channels in a token. Then each obtained attention score is normalized by a softmax operation to get A_i^s. The process can be formulated as:

$$A_i^s = \frac{\mathbf{exp}(\mathbf{W_a}^T T_i/\sqrt{d})}{\sum_{j=1}^{N} \mathbf{exp}(\mathbf{W_a}^T T_j/\sqrt{d})}, \tag{4.1}$$

where N refers to the number of tokens and $i \in [1 \dots N]$, $\mathbf{W_a} \in \mathbb{R}^d$ is learnable weights of linear transformation. The final global attention is obtained by weighted sum:

$$A^g = f(T) = \sum_{i=1}^{N} A_i^s \cdot T_i. \tag{4.2}$$

Note that additive attention has multiple heads as in the standard self-attention.

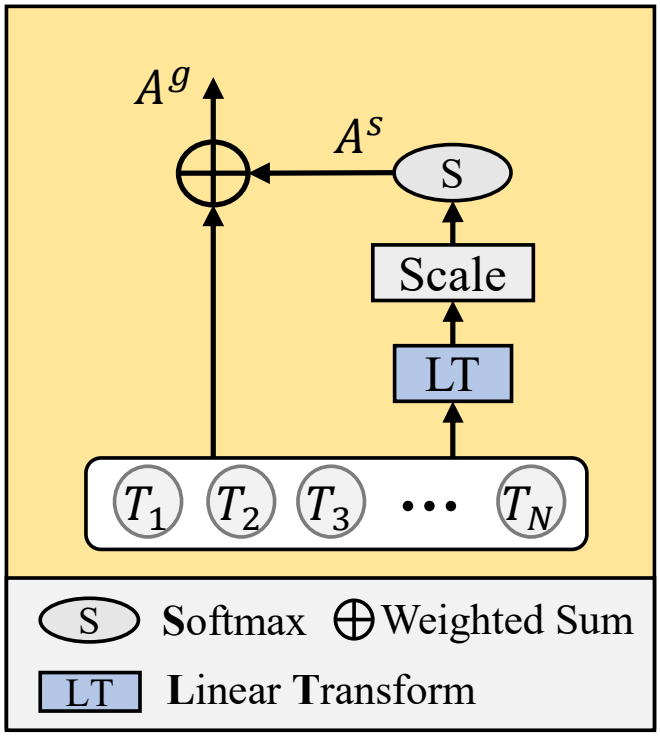


Figure 4.2: Structure of the additive attention block

4.4.3 AC Block

As shown in Figure 4.3, the **A**dditive attention **C**ross fusion block (AC) adopts a symmetrical structure. $M_i \in \mathbb{R}^{d\times N}, i \in [1,2]$ denote the inputs from two encoders. Concretely, M_1 and M_2 are first processed by four linear transformation (LT) units, respectively:

$$M_i^{t_j} = \mathbf{W_i^j}^T M_i, \tag{4.3}$$

where $\mathbf{W_i^j} \in \mathbb{R}^{d\times d}$ refers to learnable parameters in LT, $i \in [1,2]$, $j \in [1,2,3,4]$. For the left part, M_1^{t1} is fed into additive attention blocks to get a global attention score $A_1^{g1} \in \mathbb{R}^d$ and then element-wise multiplied by $M_2^{t_2}$ to integrate attention score of M_1 to the feature map of M_2. Then, in the same way, we build attention scores A_1^{g2}, and A_1^{g3} while only considering the feature map of M_1. For the right part, we use the same way to get $A_2^{g_3}$. Meantime, we also introduce the information from M_1 as an additional reference. Note that we use the knowledge from another modality to calibrate the long-range dependencies building process in the current modality. This strategy makes it easier for AC block to establish cross attention from one modality to another. This process can be formalized as:

$$A_i^{g_3} = f\Big(f\big(f(M_i^{t_1}) \otimes M_{3-i}^{t_2}\big) \otimes M_i^{t_3}\Big), \tag{4.4}$$

where $i \in [1,2]$, f denotes additive attention operation (see equation 4.2) and $\otimes$ denotes element-wise multiplication

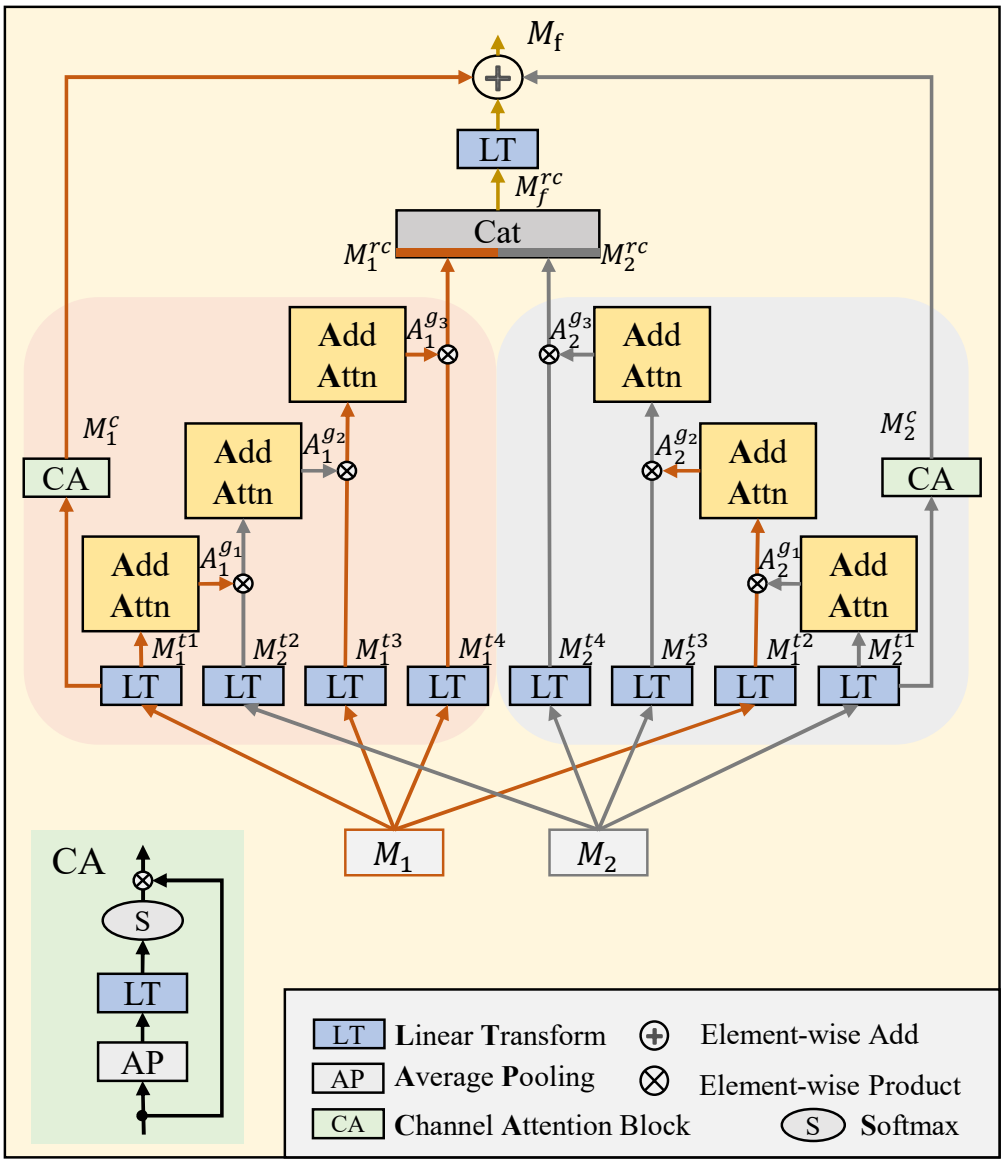


Figure 4.3: Structure of the proposed AC block in Figure. Add-Attn is the additive attention shown in Figure 4.2

operation.

At the same time, $M_1^{t_1}$ and $M_2^{t_1}$ are respectively transferred into two bypass branch modules to get M_1^c and M_2^c. The bypass branch module is designed similar to the classical channel attention (CA) mechanism, which can be described as:

$$\begin{aligned} M_i^c &= \mathbf{Softmax}(M_i^t) \otimes M_i^{t_1} \\ M_i^t &= \mathbf{W_i^c}^T M_i^a \\ M_i^a &= \mathbf{AvePooling}(M_i^{t_1}), \end{aligned} \tag{4.5}$$

where $i \in [1, 2]$, $\mathbf{W_i^c} \in \mathbb{R}^{d \times d}$ are the parameters of LT, $\mathbf{AvePooling}$ is average pooling operation along tokens.

Next, the generated $A_1^{g_3}$ and $A_2^{g_3}$ are respectively element-wise multiplied by $M_1^{t_4}$ and $M_2^{t_4}$, and then concatenated along channel axis to get M_f^{rc}. Finally, after a linear transformation, M_f^{rc} is element-wise added with M_1^c and M_2^c to get the final output M_f:

$$\begin{aligned} M_f &= M_1^c + M_2^c + M_t^{rc} \\ M_t^{rc} &= \mathbf{W^{rc}}^T M_f^{rc}, \end{aligned} \tag{4.6}$$

where $\mathbf{W^{rc}} \in \mathbb{R}^{2d \times d}$ are parameters of LT and $+$ refers to element-wise addition operation.

The idea behind this design is very intuitive. We use the global attention from M_1 to calibrate the features in M_2. Then the calibrated feature map of M_2 regenerates new global attention, which is further used to re-calibrate the features in M_1, and vice versa. Therefore, the block fully evaluates the interrelationship between different modalities to achieve a better efficient fusion. Note that AC block can also perform cross-feature fusion in sub-windows for more flexible analysis of local features.

4.4.4 EAC Block

Efficient **A**dditive attention **C**ross fusion block (EAC) is a variant of AC (section 4.4.3), designed to puzzle out the excessive consumption of building long-range dependencies under large resolution input. We consider that establishing global context information under full-resolution input will introduce redundancy, which leads to unnecessary calculations. In addition, fine-grained shape information is essential for establishing target contours. Accordingly, we decouple the process of establishing global context information and contour information. To do so, we design a shape extraction module in EAC block. An example for demonstrating the shape extraction module is shown in Figure 4.5. For each input, briefly, we compute the mean value of each local area by a sliding window with a certain stride which is equal to the size of the sliding window, to obtain the mean map. Then, the mean value in each local window is removed to obtain the shape map. Thanks to the parallel computing of Pytorch [139], this process can be implemented very efficiently.

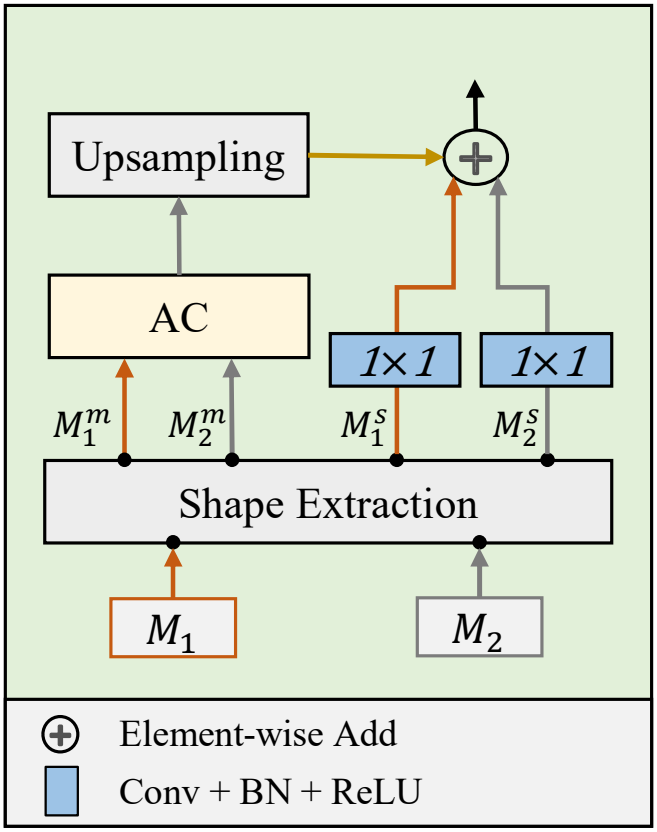


Figure 4.4: Structure of the proposed EAC block

EAC block is shown in Figure 4.4, we first build the mean map M^m and the shape map M^s of the input M_1 and M_2 through a shape extraction module. The extracted mean information M_1^m and M_2^m are input to the AC block to get the global context attention, and then followed by an upsampling operation to recover the resolution. The extracted shape information is processed by a pixel-wise convolution. Finally, mean information and shape information are integrated by element-wise addition. Table 4.1 shows the configuration of AC or EAC blocks at every encoding stage.

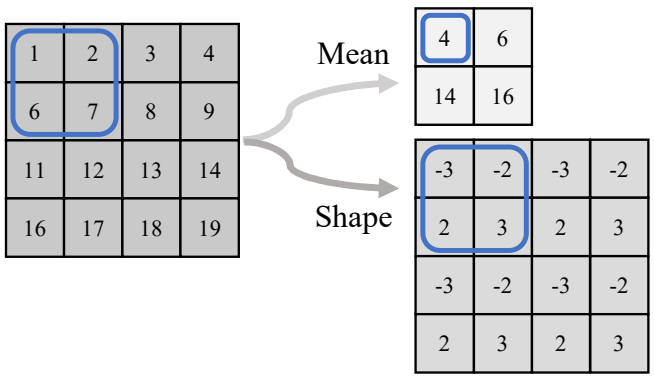


Figure 4.5: An example of Shape extraction in Figure 4.4

Table 4.1: Configuration of AC or EAC blocks in Figure 4.1

block	(c1, c2)	im_scale	sw	sub_w	heads
EAC	(64, 64)	1/2	(8, 8)	(2, 2)	8
EAC	(64, 64)	1/4	(4, 4)	(2, 2)	8
AC	(128, 128)	1/8	-	(4, 4)	16
AC	(256, 256)	1/16	-	(h/16, w/16)	32
AC	(512, 512)	1/32	-	(h/32, w/32)	64

h and w refer to the height and width of original resolution, c_1 and c_2 denote channels of each modality, im_scale denotes the ratio of the current input size to the original image size, sw denotes the size of sliding window in EAC block, sub_w denotes the size of sub-windows in AC block, and $heads$ denotes the number of head in additive attention.

4.4.5 TC Block

As shown in Figure 4.6, the **T**rans-**C**ontext block is composed of convolution and transformer blocks. Specifically, we first project the input channels through a 1×1 convolution, and then several TBs are applied to obtain complex context. Finally, we restore the number of input channels through another 1×1 convolution. The whole process is straightforward but very convenient. Note that the TB in our TC block can utilize existing well-designed methods, such as [121, 122]. We reimplement and employ TB of [127] in our TC block.

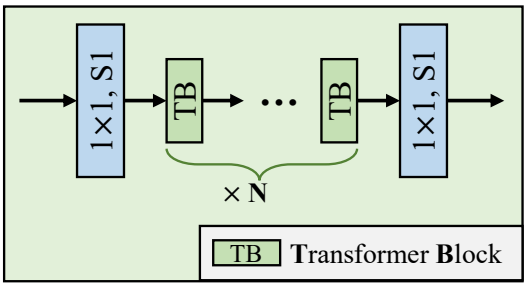


Figure 4.6: Structure of the proposed TC block in Figure 4.1

4.5 Experiments

4.5.1 Datasets

NYUv2. NYUv2 is a popular dataset for indoor scene analysis. It contains 1449 indoor finely annotated RGB-D images, in which 795 are used for training and 654 for testing. All images are provided with a resolution of 640×480. We follow [70] using the train/test splits as provided by the dataset and report results on the 13 and 40 classes [128] settings.

Cityscapes. The Cityscapes dataset is a large-scale database for urban street scene parsing. It contains 5000 finely annotated images captured from 50 cities with 19 semantic object categories, in which 2875 images are used for

training, 500 images and 1525 images are used for validation and testing separately. All images are provided with a resolution of 2048×1024. We report results on the reduced 11 classes [116] setting.

4.5.2 Implementation Details

We implement our network based on Pytorch [139], and all experiments are run on a Nvidia RTX3090 GPU with 24GB memory. For the network, we take Resnet-34 initialized with the pre-trained weight on ImageNet [6] as the backbone of both encoders. We train our model for 500 epochs with a mini-batch size of 8 for the NYUv2 dataset and 300 epochs with a mini-batch size of 16 for the Cityscapes dataset. As for optimization, NYUv2 dataset is trained on SGD optimizer with a initial learning rate of 0.015 and Cityscapes dataset is trained on Adam optimizer with an initial learning of 0.0001. Following [136], we employ a one-cycle learning rate policy. Moreover, we set the number of TB in TC block (N) as 3. The image input size is set to 640×480 on the NYUv2 dataset and 768×384 on the Cityscapes dataset. If not otherwise noted, the inputs of all models are RGB and depth images. Note that before training on the Cityscapes dataset, we follow the official guide to generate a depth map from the original disparity data [31]. Random scaling, cropping, and flipping are applied for data augmentation to increase the number of training samples further. We evaluate our model based on mean intersection over union (mIoU). In addition, we still care for the frame per second (FPS) rate because of the computational burden.

4.6 Results

4.6.1 Results on NYUv2

Table 4.2 compares the performance of our proposed methods with start-of-the-art methods. For a comprehensive comparison, we re-implement the prevalent multi-modal fusion methods based on their official repository and report the results on NYUv2-13 setting. Besides, we report our results on the commonly used NYUv2-40 setting. For the methods tested in the original paper, we use the reported results directly. We then follow [136] to modify our model by replacing BasicBlock with Non-Bottleneck-1D-Block (NBt1D) [140]. In our experiments, we also pay attention to FPS since they reflect the actual operating efficiency of the model. All FPS are executed at the input resolution of 640×480 on a laptop with Intel i7-9750 CPU and Nvidia RTX 2080 8G GPU. We noticed that the model based on NBt1D runs slower than the original model, which is inconsistent with the report in [136]. We consider this is because the 3×3 convolution is fully optimized in the computer environment. In table 4.2 we can see that our model outperforms current state-of-the-art methods on both NYUv2-13 and -40 classes settings while keeping a fast inference time. In addition, we found that our method is capable of capturing overall contextual information while extracting valuable details. Please refer to the supplementary material for some qualitative results.

Table 4.2: Performance of different methods on NYUv2 test set.

Model	BackBone	mIoU (%)		FPS
		NYUv2-13	NYUv2-40	
FuseNet[78]	Vgg-16	54.6	-	15.1
RedNet[117]	ResNet-50	64.0	-	24.2
ACNet[82]	ResNet-50	64.8	48.3*	17.9
ESANet†[136]	ResNet-34	65.1	50.3*	39.7
ESANet[136]	ResNet-50	65.9	50.5*	44.6
ShapeConv[70]	ResNext-101	65.1*	51.3*◇	12.3
HCFNet(Ours)	ResNet-34	65.8	49.9	36.2
HCFNet†(Ours)	ResNet-34	66.7	50.7	31.7
HCFNet(Ours)	ResNet-50	66.9	51.5	22.5

∗ denotes that we report the result from the original paper, † denotes that the BasicBlock is replaced by NBt1D [140], and ◇ refers to multi-scale testing strategy.

Table 4.3: Performance of different methods on Cityscapes-11 val set.

Model	BackBone	mIoU (%)	FPS	Latency
RedNet[117]	ResNet-50	79.6	26.1	0.038
ACNet[82]	ResNet-50	80.0	19.6	0.051
ESANet[136]	ResNet-34	77.8	47.6	0.021
ESANet†[136]	ResNet-34	78.5	42.1	0.024
HCFNet (Ours)	ResNet-34	78.4	39.0	0.025
HCFNet† (Ours)	ResNet-34	78.9	35.2	0.028
HCFNet (Ours)	ResNet-50	80.6	25.3	0.039

∗ denotes that we report the result from the original paper and † denotes that the BasicBlock is replaced by NBt1D [140].

4.6.2 Results on Cityscapes

To exhibit the capabilities of our model in outdoor scenarios, we evaluate our model on the Cityscapes-11 dataset. Specifically, we resize the input image to a resolution of 768×384. All models are trained and evaluated at the same resolution. Note that all models are configured to the same training strategy, unless a different setting is provided in the original implementation. We observed that the Cityscapes dataset is very sensitive to the backbone, and a well pre-trained backbone can significantly improve the performance. As shown in Table 4.3, our model yielded a very comparable result when using ResNet34 as backbone, and improved the segmentation results when using ResNet50 as backbone.

4.6.3 Ablation Analysis

To verify the functionality of the components of our model, we conduct an ablation study on the NYUv2-13 dataset. We use the network architecture in Figure 4.1 as the basic structure. For a fair comparison, network architecture and hyper-parameters in different experiments are fixed. In ablations, we first evaluate the influence of different fusion methods, namely Add, ACM, SE, which have been used in recent start-of-the-art models. Specifically, Add [78, 117] simply adds the data of different modalities directly. ACM [82] and SE [136] calculate each modality's attention through the Attention Complementary module (ACM) and Squeeze-and-Excitation module (SE) before fusion. Then we estimate the impact of the TC block on performance. Table 4.4 summarizes the results of this ablation study on HCFNet.

From Experiment 1 to Experiment 3, we use different fusion methods to replace AC and EAC blocks of the original HCFNet while removing the TC block. As we can see, the proposed AC method is superior to other fusion methods by a large margin. Figure 4.7 visualized the feature maps of the output of different fusion methods at different stages in HCFNet. Specifically, in B1 and B2, we use the EAC blocks. We can see that EAC can remove redundant facts in the scene without losing valuable information. In B3-B5, AC blocks are deployed. It can be seen that the targets of interest are effectively activated, and it has a more extensive range and more accurate position than other fusion methods. This further verifies that the global awareness obtained in the AC module helps the model understand the scene. Please refer to the supplementary material for additional analysis of the AC/EAC block.

In addition, after deploying the TC block, the performance of our model is significantly improved (+0.6%), which reveals that it is practical to further establish long-range dependencies in the fused features. In the final experiment, we replaced the EAC block in Table 4.1 with the AC block, which caused a decrease in model performance. This echoes our previous argument that the context contains much redundant information under large-resolution input, which will lead to unnecessary calculations and confusion. In contrast, our proposed AC block can efficiently establish long-range awareness, and with the help of the EAC block, redundant information can be eliminated without destroying local details while significantly reducing the amount of calculation. Finally, the proposed TC block can further coordinate information fusion and establish a more significant receptive field.

4.7 Summary

In this chapter, we delve into the exploration of harnessing depth images to provide additional geometric cues under a multi-modal framework. The aim is to enhance the perceptual abilities of semantic segmentation models, in which we employ cross-attention mechanisms to selectively focus on salient features from one modality while downplaying less informative ones from another.

Specifically, we design a novel multi-modal visual data fusion method, which can efficiently integrate data from

Table 4.4: Comparison of different fusion methods and components.

	Model	Fusion	Context	mIoU (%)
1	HCFNet	Add[117]	None	63.3
2		ACM[82]	None	64.1
3		SE[136]	None	63.9
4		AC* (Ours)	None	65.2
5		AC* (Ours)	TC	65.8
6		AC (Ours)	TC	65.1

⋆ denotes we follow Table 4.1 to configure AC and EAC blocks.

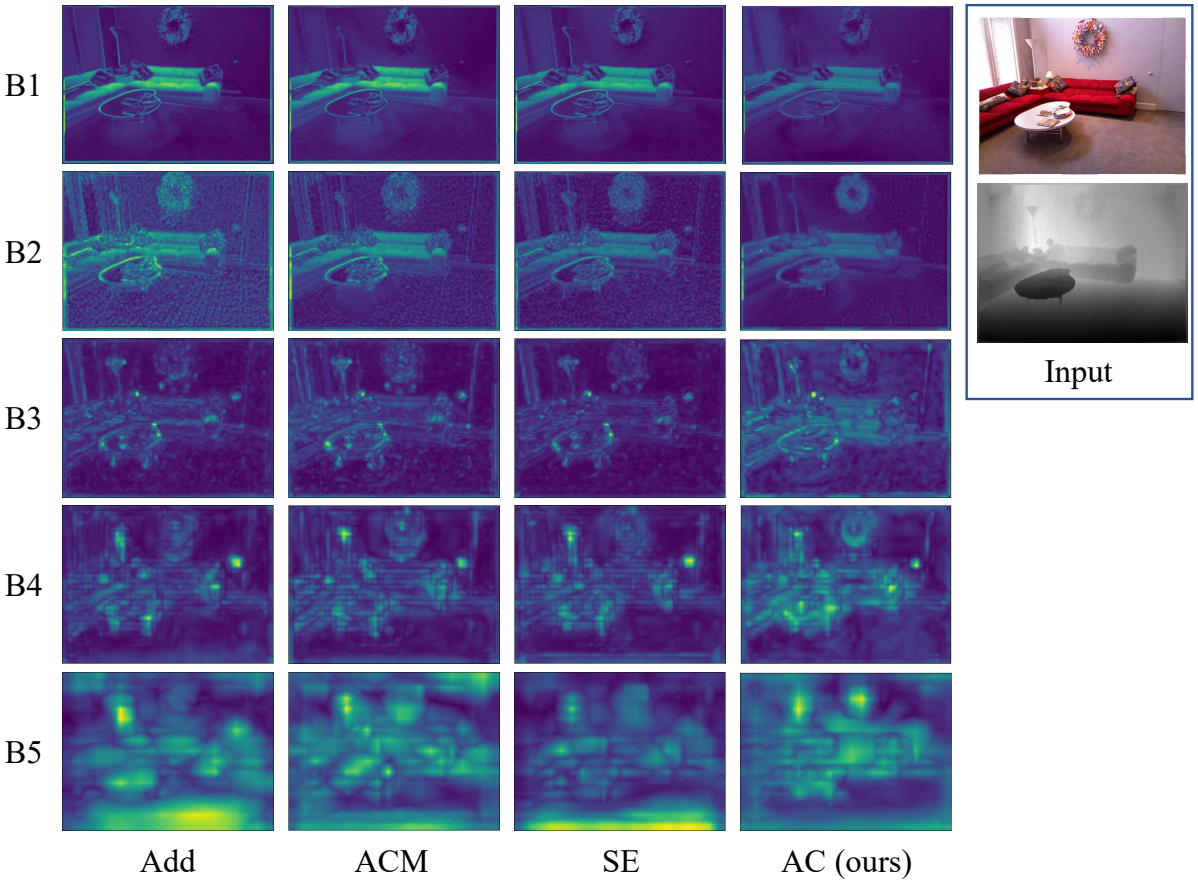


Figure 4.7: Visualization of feature maps of different fusion methods. B1-B5 refers to the output of different fusion blocks in the encoding part of Figure 4.1. Note that the sample comes from the NYUv2 test set, and all outputs are resized to a resolution of 640×480 for a best of view.

different modalities. It also ensures that the model retains valuable local details after fusion while having a global receptive field. Precisely, we customize a multi-modal fusion block named AC block based on the additive attention mechanism, which assists in forming global awareness inter- and inner-modalities. Then, we propose the EAC block, an efficient variant of the AC block, to efficiently build global attention and keep details under high-resolution input. On the other hand, based on the transformer block, we offer a simple yet effective context fusion block called trans-context (TC) block for further connecting the context output from the encoder. Together with the proposed well-designed components, we present HCFNet for semantic segmentation of indoor and outdoor scenarios. Finally, comprehensive experiments and ablation studies verify the effectiveness of our network and different components.

Chapter 5

Channel-Patch Cross feature fusion for RGB-T Object Detection

We have already established that complementary information from various modalities can enhance the perception capability of a system to its environment. Chapter 3 and Chapter 4 have explored methods for fusing RGB and Depth images, effectively incorporating spatial context and depth information to create a more holistic understanding. However, some sensors might fail to function under adverse weather conditions. For instance, traditional visible light sensors struggle to capture sufficient information to accurately represent the environment in low-light conditions. On the other hand, thermal infrared sensing, with its consistent imaging performance full-time, is widely used in low-light conditions, such as at night. Therefore, effectively combining thermal infrared images with RGB images can enhance the system's accuracy and robustness under varying lighting conditions. Nevertheless, unlike RGB-D data fusion, where RGB acts as the primary signal and depth serves as an auxiliary, the fusion of RGB and thermal infrared images is more flexible. Under low-light conditions, thermal infrared images contain more valuable information, while RGB images offer more abundant background and texture details under ample illumination. Given this context, at the core of fusion lies how to mutually rectify features across different modalities rather than merely using one modality as a supplement to another. To this end, this chapter is devoted to designing an effective RGB-T fusion method. While exploring the complementarity of different modalities, it also promotes the inter-calibration of various modal features, thereby allowing the system to form a more comprehensive and nuanced interpretation of the surrounding context.

5.1 Abstract

Data from different modalities, such as infrared and visible light images, can offer complementary information, and integrating such information can significantly enhance the perceptual capabilities of a system to the surroundings. Thus, multi-modal object detection has widespread applications, particularly in challenging weather conditions like low-light scenarios. The core of multi-modal fusion lies in developing a reasonable fusion strategy, which can fully exploit the complementary features of different modalities while preventing a significant increase in model complexity. To this end, this chapter proposes a novel lightweight cross-fusion module named Channel-Patch Cross Fusion (CPCF), which leverages Channel-wise Cross-Attention (CCA) and Patch-wise Cross-Attention (PCA) to encourage mutual rectification among different modalities. This process simultaneously explores commonalities across modalities while maintaining the uniqueness of each modality. Furthermore, we design a versatile intermediate fusion framework that can leverage CPCF to enhance the performance of multi-modal target detection. The proposed method is extensively evaluated on multiple public multi-modal datasets, namely FLIR, LLVIP, and DroneVehicle. The experiments indicate that our method yields consistent performance gains across various benchmarks and can be extended to different types of detectors, further demonstrating its robustness and generalizability. Our codes are available at http://github.com/

5.2 Introduction

Object detection involves extracting items of interest from input data and locating their positions, which has a wide range of applications in the real world, such as autonomous driving [141], security surveillance [142], and disaster relief [143]. In recent years, numerous advanced object detection methods have emerged [63, 52, 62], demonstrating outstanding performance on various tasks with color images, i.e., RGB, as inputs [7, 64]. However, real-world scenarios are often dynamically changing, and it is impossible to gather sufficient clues to detect all objects in a scene merely through the color modality. For instance, the image quality captured by visible light cameras at night typically deteriorates significantly, substantially reducing the accuracy and robustness of detection results.

On the other hand, due to the stability in imaging under different lighting conditions, thermal infrared cameras are frequently employed in low-light situations to enhance the system's capability to capture scene information. More concretely, thermal images are used to provide full-time geometric characteristics of objects, such as shape and contour, while color images provide rich texture information when light is sufficient. Therefore, an effective fusion strategy is needed to fully exploit the complementary features among these different modalities. In this context, several studies [144] seek to leverage different fusion strategies to explore the optimal joint representation of RGB and thermal images, which in turn improves the model's capability to perceive the surroundings under low-light conditions. According to the location of fusion occurrence, multi-modal fusion can be categorized into early

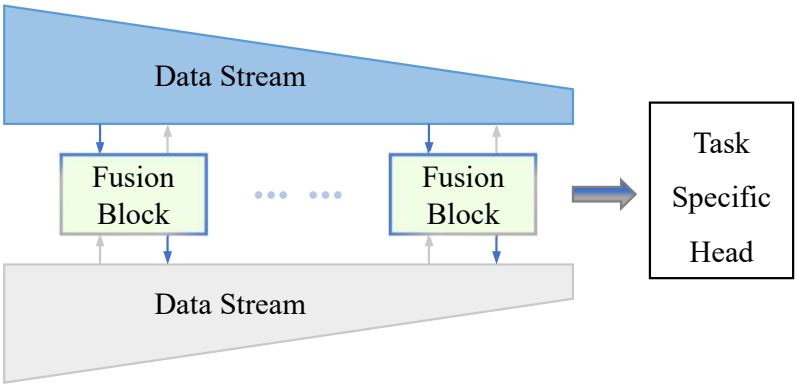


Figure 5.1: Framework of intermediate multi-modal visual data fusion.

fusion, late fusion, and intermediate fusion as discussed in chapter 2, section 2.3.

Specifically, early fusion directly concatenates multi-modal data into a unified multi-channel input, which is then fed into a general object detection network. Conversely, late fusion independently processes data from different modalities and integrates the outputs at the point of decision-making by an additional fusion operation. Recent studies [145, 84] have revealed some limitations of early fusion and late fusion, such as early fusion struggles to effectively integrate specific modality features while late fusion suffers from a lack of feature interaction between different modalities. Therefore, between early and late fusion, intermediate fusion incrementally merges features of different modalities through a flexible structure design, allowing the features to maintain their independence while interacting. The typical intermediate fusion framework is illustrated in Figure 5.1, where multi-modal data is processed through two feature extraction networks, known as backbones, to refine crucial features. Meanwhile, the fusion module integrates multi-modal information and redistributes it back into the original data streams. Although intermediate fusion presents advantages, the design of efficient fusion modules to accurately integrate diverse features and maintain the integrity of original data still poses a significant challenge. To this end, some works [146, 147, 148] have attempted to dig latent relationships among different modalities through attention mechanisms and achieved promising results. In addition, self-attention [108, 149] has been shown to be an effective way for establishing long-range connections, which can effectively leverage the complementary characteristics between different modalities by constructing associations among their contexts. Yet, the extensive computation required by attention mechanisms significantly constrains its potential in multi-modal fusion.

We argue that at the core of fusion lies the question: "*how to mutually rectify features across different modalities rather than merely using one modality as a supplement to another*?". Thereby, this chapter focuses on harnessing self-attention to fully explore the inherent complementarity between different modalities to facilitate the efficiency of mutual fusion and rectification. To this end, we propose a lightweight cross-attention fusion module, termed channel-patch cross fusion (CPCF), which is composed of channel-wise cross-attention (CCA) and patch-wise cross-attention (PCA). Specifically, we employ parametric-free operations such as average pooling and max pooling to

model the characteristics of each modality and incorporate cross-attention to reconstruct complementary awareness across different modalities in terms of channels and spatial dimensions, thus ensuring the complementarity of different modalities while maintaining their independence. Note that the features of different modalities are compressed by basic operations such as pooling, which makes the additional complexity generated by the following cross-attention can be negligible. Consequently, it allows the module to significantly enhance the performance of multi-modal object detection while minimizing the impact of cross-attention on computational efficiency. To demonstrate the efficacy of CPCF, we design a general intermediate fusion architecture, as depicted in Figure 5.2, which can be extended to various detectors. Besides, we conduct extensive experiments on the generic multispectral dataset named FLIR [150] and LLVIP [151] and a more challenging oriented object detection dataset called DroneVehicle [146]. Our results demonstrate that our proposed approach can remarkably improve the performance of object detection without significantly increasing the complexity of the model.

The contributions of this chapter are summarized as follows:

- We propose a lightweight channel-patch cross fusion (CPCF) module to construct cross-modal features in both channel and spatial dimensions, during which the CPCF module leverages the properties specific to one modality to calibrate the features of another, thus effectively modeling the complementary properties between modalities and optimizing the representability of features in the data stream.
- We design an intermediate fusion framework based on CPCF, which can be flexibly integrated into various object detection frameworks to efficiently exploit multi-modal cues to boost the performance of models.
- We conduct extensive experiments on different types of multi-modal datasets and obtain optimal results. Simultaneously, we validate the generalization ability of our method on different detectors, which further shows its robustness and versatility.

The rest of this chapter is organized as follows: Section 5.3 reviews the existing works related to our method. The overall framework is presented in section 5.4 with the details of channel-wise cross-attention and patch-wise cross-attention. The experimental setup and the results are presented and discussed in section 5.5. Finally, Section 5.6 ends this paper with a conclusion and discussion.

5.3 Related Work

5.3.1 Unimodal Object Detection

Unimodal object detection typically employs RGB images as input, which can be categorized into two- and single-stage approaches. Two-stage approaches divide object detection into two distinct phases, i.e., the regional proposal phase and the target classification and bounding box regression phases. As a trailblazing effort, RCNN [50] leverages

the selective search algorithm [51] to generate numerous potential regions, then employs SVM and a regressor for classification and bounding box prediction tasks. Next, FastRCNN [11] and FasterRCNN [52] upgrade this idea within a deep learning framework, further improving training efficiency and model performance. On the other hand, single-stage object detection frameworks, represented by YOLO [54], directly predict the category and location of objects in a single forward propagation, eliminating the need for a region proposal stage, thereby greatly enhancing the detection speed. Especially some variants [56, 57, 60] of YOLO are gradually catching up with the two-stage detector in terms of detection accuracy while maintaining high operating speed. Recently, YOLOX [63] has transformed the YOLO detector into an anchor-free style, further enhancing processing speed. Meanwhile, it capitalizes on strong data augmentation and advanced label assignment strategies for superior performance.

Moreover, in some special scenarios, such as remote sensing images, traditional axis-aligned bounding boxes cannot accurately describe the state of objects. For this reason, oriented object detectors [65, 67] are designed to align the bounding boxes with the orientation of the targets. These detectors rely on existing object detection frameworks and predict the direction of bounding boxes through additional modules. For instance, S^2A-Net [66] introduces a feature alignment module and an oriented detection module for mitigating the misalignment between oriented anchors and axis-aligned convolutional features. Then, the PSC [152] utilizes an additional phase shift encoder to achieve an accurate prediction of the orientation.

In this work, we implement our method within different detectors and conduct extensive experiments on different types of datasets to tackle multi-modal object detection tasks under various scenarios.

5.3.2 Multi-modal Object Detection

Multi-modal object detection, which merges various types of data to bolster the robustness and accuracy of object detection tasks, is a vibrant research field in the computer vision community. It typically blends multi-modal data through early fusion, late fusion, or intermediate fusion strategies. In early fusion, RGB and IR images are concatenated at pixel level to form a 4-channel input, and then features are extracted with a regular object detection framework. However, early fusion forgets modality-specific properties during feature forward propagation, which can lead to suboptimal results [72, 153]. Conversely, late fusion process each modality independently through separate models and the results are merged at the decision level [154]. For instance, ProbEn [155] utilizes the Bayesian rule to find the optimal fusion strategy to ensemble the results of RGB and IR data streams. However, assembling multiple detectors results in more false positive cases and slower detection speeds [148]. On the other hand, intermediate fusion lies between the two, in which features from different modalities interact with each other while still preserving their individuality [156, 146, 157, 158, 148]. For instance, GFD-SSD [156] employs two encoders to handle RGB and thermal images and utilizes a gating unit to merge features from the intermediate layers in a single-stage detector framework. UA-CMDet [146] proposes an uncertainty-aware module to reduce the detection bias caused

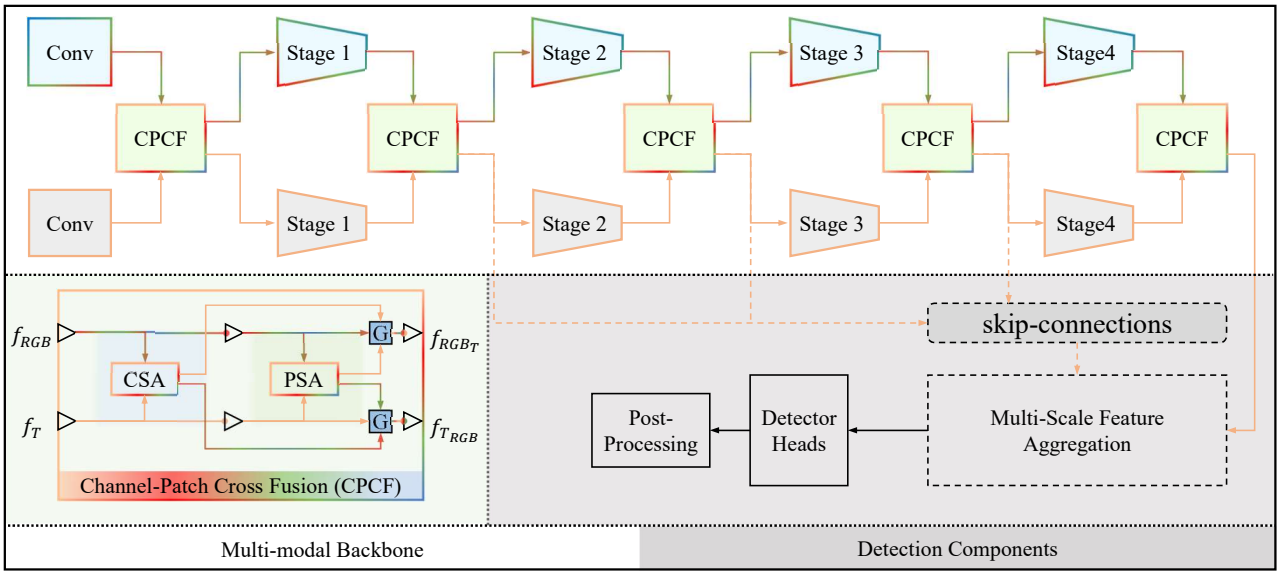


Figure 5.2: Overview of CPCF-based Object Detection Framework.

by high-uncertainty objects. Moreover, inspired by the attention mechanism, GAFF [157] and ECISNet [158] leverage spatial attention to learn the adaptive weighting and fusion of different modalities. Then, BAANet [159] design a bi-directional adaptive attention gate to recalibrate and fuse multi-modal information in both channel and spatial dimensions. Recently, CMAFF [147] proposes a lightweight attention module to extract shared features across modalities while emphasizing inter-modal differences. At the same time, CSAA [148] utilizes a channel switching strategy in the attention module which also reduce the computational conplexity of the fusion process.

In this work, inspired by self-attention [108], we employ parameter-free operations to condense features and calculate cross-modal attention separately from both channel and spatial dimensions, which significantly reduces the computational load of CPCF. Besides, we leverage the learnable gating units to adaptively integrate different attention at different levels rather than treating them equally.

5.4 Method

In this section, we propose a lightweight cross-fusion module named CPCF, which can efficiently build long-range dependencies from one modality to another in both channel and spatial axes. Building upon CPCF, we further design a generalized intermediate fusion object detection framework to effectively exploit multi-modal information. In the following, we will detail the proposed intermediate fusion framework and the associated modules.

5.4.1 Framework Overview

As shown in Figure 5.2, the overall multi-modal object detection framework is composed of two part. The first part is a general multi-modal backbone, an intermediate fusion-based feature extractor for refining and fusing multi-

modal information. The second part is the detection related components, which provide modules, such as skip connections and detection heads, for different types of detectors. Basically, the multi-modal backbone originates from prevalent single-modality backbones, such as ResNet [9] and CSPDarknet [160], which are typically composed of several convolution stages, enabling a more efficient and comprehensive encoding of information from inputs. As illustrated in the upper half of Figure 5.2, we employ a symmetrical structure to separately process information from different modalities. Meanwhile, the proposed CPCF module is deployed subsequent to each convolution stage to calculate the cross-attention across different modalities and recalibrating the features accordingly. Afterward, the calibrated features are propagated to the components specific to the object detection tasks, as shown in the lower right of Figure 5.2. Taking YOLOX [63] as an example, the fused features from different CPCF modules are aggregated via a feature pyramid module to multiple object detection heads for multi-scale prediction. In addition, for the two-stage detector, like RCNN [50], a region proposal module is operated to receive the outputs from the last CPCF module.

5.4.2 Multi-modal Cross-Attention

While different visual modalities carry complementary information valuable for perception tasks, they also contain a considerable amount of redundant data and noise, factors that can potentially influence the efficiency of data analysis and interpretation. In this context, we propose a multi-modal cross-attention mechanism that calibrates one modality with the features of another. This structure amplifies the complementary characteristics between modalities while diminishing redundant information, thereby fostering a more effective and integrated multi-modal representation. Basically, the feature representation of a modality can be reflected in both channel and spatial dimensions. Accordingly, we create channel-wise cross-attention (CCA) and patch-wise cross-attention modules (PCA) among different modalities, in which CCA is designed to establish channel relationships, and PCA is expected to build spatial relationships during the feature extraction process. This strategic design enables cross-modality feature recalibration, ensuring a more cohesive and effective multi-modal data integration.

5.4.2.1 Channel-wise Cross-Attention

In a feature map, a channel is usually treated as a feature detector [106], thus channel-wise cross-attention (CCA) is designed to highlight beneficial channels across different modalities and suppress noise-included ones. To this purpose, CCA considers feature channels of two modalities parallelly and associates different attention weights to different channels. The overall architecture of CCA is shwon in Figure 5.3.

Specifically, given the intermediate feature maps $f_{RGB} \in \mathbb{R}^{C\times H\times W}$ and $f_T \in \mathbb{R}^{C\times H\times W}$ of two modalities, average-pooling (AP) and max-pooling (MP) are applied to compress spatial information, followed by a series of subtraction operation to obtain the cross-modal differential signals. These are then concatenated into compact

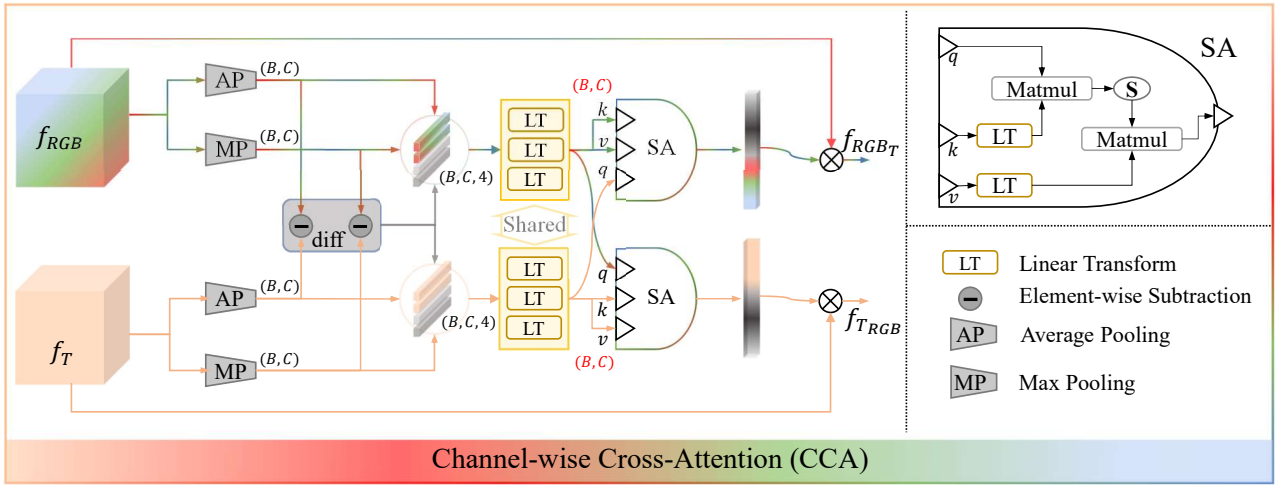


Figure 5.3: Channel-wise Cross-Attention.

expressions $f_{RGB}^{C} \in \mathbb{R}^{C\times 4}$ and $f_{T}^{C} \in \mathbb{R}^{C\times 4}$, as expressed as follows:

$$
\begin{aligned}
f_{diff}^{AP} &= |\mathbf{AP}(f_{RGB}) - \mathbf{AP}(f_T)|, \\
f_{diff}^{MP} &= |\mathbf{MP}(f_{RGB}) - \mathbf{MP}(f_T)|, \\
f_{RGB}^{C} &= Concat\left([\mathbf{AP}(f_{RGB}), \mathbf{MP}(f_{RGB}), f_{diff}^{AP}, f_{diff}^{AP}]\right), \\
f_{T}^{C} &= Concat\left([\mathbf{AP}(f_T), \mathbf{MP}(f_T), -f_{diff}^{AP}, -f_{diff}^{AP}]\right).
\end{aligned}
\tag{5.1}
$$

Inspired by traditional self-attention [108], we seek to construct long-range dependencies of each channel. Basically, the original self-attention encodes the inputs into a set of vectors, i.e., Query (Q), Key (K), and Value (V). Then, the self-attention map is computed via a matrix multiplication QK^T. After that, the output of self-attention is obtained by another matrix multiplication between the attention map and V, which can be described as follows:

$$
f^{SA} = \mathbf{SA}\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{D_k}}\right)V. \tag{5.2}
$$

where $\frac{1}{\sqrt{D_K}}$ is a scaling factor. In this manner, the module can construct global attention across tokens.

In contrast, the compressed channel features are expressed in vector form, making them inherently compatible with self-attention. To be specific, we regard each channel as a token and project them into vectors designated as $Q \in \mathbb{R}^{C\times 1}$, $K \in \mathbb{R}^{C\times 1}$, and $V \in \mathbb{R}^{C\times 1}$ through a straightforward linear transformation:

$$
\begin{aligned}
Q &= f_X^C W_Q, \\
K &= f_X^C W_K, \\
V &= f_X^C W_V,
\end{aligned}
\tag{5.3}
$$

where $W_Q \in \mathbb{R}^{4\times 1}$, $W_K \in \mathbb{R}^{4\times 1}$, and $W_V \in \mathbb{R}^{4\times 1}$ are weight matrices of linear transformation, and the subscript X

is either RGB or Thermal. When computing self-attention, we swap the Q vector of the two modalities rather than directly using them for the attention calculation, thus forming cross-attention. As illustrated in the SA module in Figure 5.3, considering that the computational cost of self-attention is quadratic to the vector length, two linear transformations are employed to compress vectors K and V to reduce the computational burden. The cross-attention scores $S_{RGB}^{CA} \in \mathbb{R}^{C\times 1}$ and $S_{T}^{CA} \in \mathbb{R}^{C\times 1}$ can be formulated as follows:

$$\begin{aligned} S_{RGB}^{CA} &= \mathbf{SA}(Q_T, K_{RGB}, V_{RGB}), \\ S_{T}^{CA} &= \mathbf{SA}(Q_{RGB}, K_T, V_T). \end{aligned} \tag{5.4}$$

Finally, the attention scores from different modalities are normalized to the range $[0, 1]$ through a sigmoid function, and the channel-wise recalibrated features f_{RGB}^{RC} and f_{T}^{RC} can be described as:

$$\begin{aligned} f_{RGB}^{RC} &= \sigma\left(S_{RGB}^{CA}\right) \otimes f_{RGB}, \\ f_{T}^{RC} &= \sigma\left(S_{T}^{CA}\right) \otimes f_{T}, \end{aligned} \tag{5.5}$$

where $\sigma(\cdot)$ indicates the sigmoid function, and $\otimes$ indicates element-wise multiplication.

5.4.2.2 Patch-wise Cross-Attention

Contrary to the aforementioned CCA, which attempts to establish long-range attention across channels, patch-wise cross-attention (PCA) aims to model inter-patch connections of different modalities and leverage this to calibrate the multi-modal features across spatial. To achieve this goal, given the intermediate feature maps $f_{RGB} \in \mathbb{R}^{C\times H\times W}$ and $f_T \in \mathbb{R}^{C\times H\times W}$ and patch size $h \times w$, we first apply patch average pooling (PAP) and patch max pooling operations (PMP) to condense local information and reduce the spatial resolution of the features. Then, following the same approach as described in Section 5.4.2.1, we obtain compact expressions along the spatial dimension. The procedure can be precisely described as:

$$\begin{aligned} f_{diff}^{PAP} &= |\mathbf{PAP}(f_{RGB}) - \mathbf{PAP}(f_T)|, \\ f_{diff}^{PMP} &= |\mathbf{PMP}(f_{RGB}) - \mathbf{PMP}(f_T)|, \\ f_{RGB}^{C} &= Concat\left([\mathbf{PAP}(f_{RGB}), \mathbf{PMP}(f_{RGB}), f_{diff}^{PAP}, f_{diff}^{PAP}]\right), \\ f_{T}^{C} &= Concat\left([\mathbf{PAP}(f_{T}), \mathbf{PMP}(f_{T}), -f_{diff}^{PAP}, -f_{diff}^{PAP}]\right), \end{aligned} \tag{5.6}$$

where $f_{RGB}^{C} \in \mathbb{R}^{C\times N\times 4}$ and $f_{T}^{C} \in \mathbb{R}^{C\times N\times 4}$ denote the compact RGB and Thermal features, $N = hw$ denotes patch numbers.

Next, we utilize two separate linear transformation blocks to encode f_{RGB}^{C} and f_{T}^{C} into their corresponding $Q \in \mathbb{R}^{N\times C}$, $K \in \mathbb{R}^{N\times C}$, and $V \in \mathbb{R}^{N\times C}$ vectors. We can then compute the cross-attention scores $S_{RGB}^{CA} \in \mathbb{R}^{N\times 1}$ and $S_{T}^{CA} \in \mathbb{R}^{N\times 1}$ using Equation 5.4. Finally, the patch-wise recalibrated features f_{RGB}^{RP} and f_{T}^{RP} can be formulated

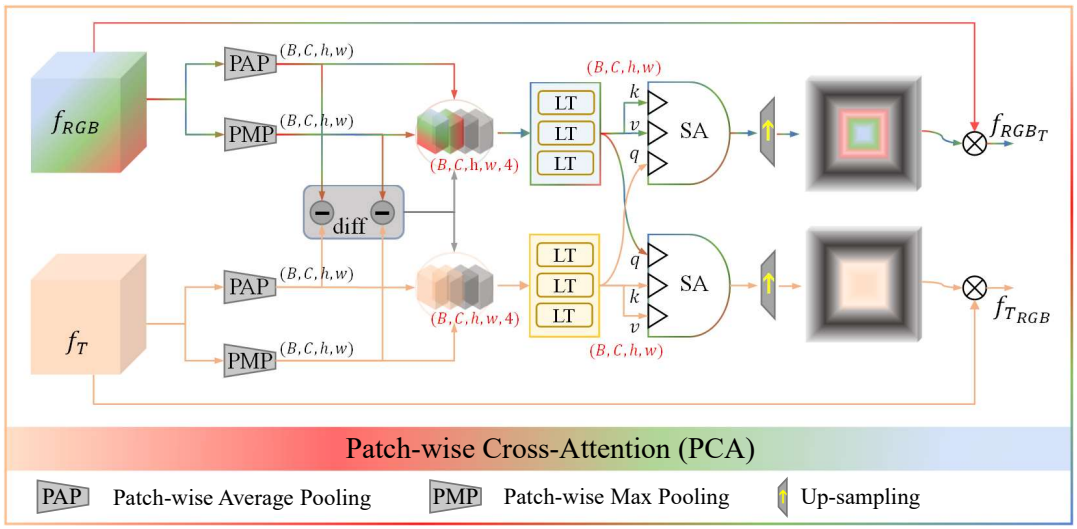


Figure 5.4: Patch-wise Cross-Attention.

as:

$$
\begin{aligned}
f_{RGB}^{RP} &= \sigma\left(UPS\left(S_{RGB}^{CA}\right)\right) \otimes f_{RGB}, \\
f_{T}^{RP} &= \sigma\left(UPS\left(S_{T}^{CA}\right)\right) \otimes f_{T},
\end{aligned}
\tag{5.7}
$$

where $UPS(\cdot)$ denotes up-sampling the size of attention scores to the input resolution. The details of PCA are depicted in Figure 5.4.

5.4.3 Channel-Patch Cross Fusion

The architecture of channel-patch cross fusion (CPCF) is shown in the lower left of Figure 5.2. In CPCF, we integrate the proposed CCA and PCA into the fusion process, thus allowing for the effective utilization of multi-modal cues and enhancing the representative capability of the fused features, further drawing out valuable information. However, treating channel and spatial attention equally during this process may lead to suboptimal results. The feature extraction process is characterized by the continuous compression of spatial resolution and expansion of channel dimensions. Throughout this process, the quantity of information across different dimensions does not remain constant.

In response to this situation, we design an adaptive gating (AG) strategy that dynamically allocates weights to different attention mechanisms, which allows a more responsive and adaptive fusion. More specifically, two learnable scaling factors, denoted as α_1 and α_2, are defined to dynamically adjust the weights of CCA and PCA during training. Then, the corresponding weights s_1 and s_2 can be computed as:

Table 5.1: Dataset Setup.

Setup	FLIR	LLVIP	DroneVehicle
Class Num.	3	1	5
Modality	RGB&Thermal	RGB&IR	RGB&IR
Box Type	Horizontal Box	Horizontal Box	Oriented Box
Img Size (original)	640×512	1280×1024	640×512
Img Size (train)	640×512	640×512	640×512
Epoches	13	13	36
Leraning Rate	2e-3	2e-3	2.5e-3
Batch Size	8	8	2
Train/Val/Test (pairs)	4139/1013/-	12025/-/3463	17990/1469/8980

$$
\begin{aligned}
s_1 &= \frac{\sigma(\alpha_1/T)}{\sigma(\alpha_1/T) + \sigma(\alpha_2/T)}, \\
s_2 &= \frac{\sigma(\alpha_2/T)}{\sigma(\alpha_1/T) + \sigma(\alpha_2/T)},
\end{aligned}
\tag{5.8}
$$

where the $\sigma(\cdot)$ denotes the sigmoid function, T is a temperature coefficient used to smooth the scaling weights. In short, given the input feature maps f_{RGB} and f_T and recalibrated feature maps f_{RGB}^{CR}, f_T^{CR}, f_{RGB}^{PR}, and f_T^{PR}, the formulation of the fused feature can be summarized as:

$$
\begin{aligned}
f_{RGB}^{Fuse} &= f_{RGB} + s_1 \cdot f_T^{CR} + s_2 \cdot f_T^{PR}, \\
f_T^{Fuse} &= f_T + s_1 \cdot f_{RGB}^{CR} + s_2 \cdot f_{RGB}^{PR}.
\end{aligned}
\tag{5.9}
$$

5.5 Experiments

In this section, we initially perform experiments on the general-purpose object detection benchmarks, specifically FLIR [150] and LLVIP [151], to assess the efficacy of our proposed methods. Subsequently, we extend our testing to a more challenging DroneVehicle [146] dataset, which targets oriented object detection. Finally, we illustrate a series of studies to ablate different components and analyze the effectiveness of our designs.

5.5.1 Datasets

FLIR The FLIR dataset is a benchmark extensively used for evaluating multi-modal object detection, comprising a substantial number of paired RGB and thermal infrared images. In our experiments, we utilize the aligned-FLIR dataset [150], wherein RGB-Thermal image pairs are correctly aligned. This dataset features 5142 RGB-Thermal image pairs, spanning three object categories: 'person', 'car', and 'bicycle', gathered from daytime to nighttime. Among these, 4139 pairs are for training, while the remaining 1013 pairs are allocated for testing.

LLVIP The LLVIP [151] is a recently introduced, large-scale dataset explicitly designed for pedestrian detection in visible-infrared contexts. It contains 15488 image pairs, with 12,025 pairs for training and 3,463 pairs for testing. A notable characteristic of this dataset is that a majority of the images are captured under extremely low light conditions. Furthermore, all images within the dataset are stringently aligned in terms of time and space.

DroneVehicle The DroneVehicle dataset [146] is a new released multi-modal benchmark specifically designed for oriented vehicle detection from a drone's perspective. It encompasses five distinct vehicle categories, namely 'car', 'truck', 'bus', 'van', and 'freight car'. This dataset comprises 28,439 RGB-Infrared image pairs that capture a variety of settings, including urban roads, residential areas, and parking lots, from day to night with a resolution of 640$\times$512. The dataset is composed of 17,990 image pairs for training, 1,469 for validation, and 8980 pairs reserved for testing.

5.5.2 Implementation Details

Utilizing the proposed CPCF, we design an intermediate fusion architecture that can be seamlessly integrated into a range of object detection frameworks. For the practical implementation, we build our model based on a popular object detection codebase MMDetection [161], and train our models on a single NVIDIA RTX3090 GPU. In all experiments, we initialize the backbone networks using the weights pre-trained on COCO [7] for general-purpose object detection. For oriented object detection tasks, the backbone networks are initialized with weights pre-trained on the ImageNet [6]. To train the models, we employ the SGD optimizer with an initial learning rate of 2e-3 and a momentum of 0.9. For data augmentation, we apply random flipping and scale the images to a resolution of 640$\times$512. In the case of the FLIR dataset, we additionally leverage the Mosaic data augmentation technique [63] to further enrich the data for methods within the YOLO family. Subsequently, all models are trained in 13 epochs with a batch size of 8. For the DroneVehicle dataset, we set the batch size to 2 and train the model for 36 epochs. The setups of different datasets are shown in Table 5.1. For all experiments, the hyper-parameter T mentioned in Equation 5.8 is set to 1.0.

Baselines To comprehensively evaluate our method, we first implement two fusion strategies, namely Concatenate and Multi-level Sum (MLSum), for multi-modal data fusion. Specifically, Concatenate means that we concatenate RGB and Thermal images along the channel dimension, exemplifying an early fusion method. While MLSum represents an intermediate fusion method, maintaining the same structure as depicted in Figure 5.2, we substitute the CPCF with a summation operation at each stage. Furthermore, we take into account detectors that utilize either RGB or Thermal inputs, serving as uni-modal baselines for comparison.

Note that our design pertains only to the encoding part of the model, which allows us to evaluate our method across various detectors such as Fcos [62], YOLOX [63], and S^2A-Net [66]. For each detector, we conduct experiments based on the aforementioned baselines to assess the generalization capability of our proposals.

5.5.3 Evaluation Metrics

In our evaluation, we adopt three standard COCO metrics, namely mean Average Precision (mAP), mAP_{50}, and mAP_{75}, to quantify the effectiveness of the proposed method. During this process, the Intersection over Union (IoU) is employed as a criterion to classify positive and negative samples. More concretely, a detected instance is deemed a positive sample only when the IoU between the predicted bounding box and the ground truth bounding box surpasses a designated threshold, denoted as τ. Consequently, for mAP_{50} and mAP_{75}, the threshold τ is set at 0.5 and 0.75, respectively. The mAP, on the other hand, is computed with the threshold τ ranging from 0.5 to 0.95 in increments of 0.05.

5.5.4 Comparative Studies

5.5.4.1 Quantitative Results

We compare the proposed fusion methods with our baselines and other state-of-the-art methods on FLIR, LLVIP, and DroneVehicle datasets. The experimental results on the FLIR dataset are illustrated in Table 5.2. Among them, Fcos [62], YOLOv5 [162], and YOLOX [162] are initially designed for RGB-based object detection, while GAFF [157], CFT [149], YOLOFusion [147], and UA-CMDet [146] are multi-modal-based object detection methods. Then, we extend the unimodal methods to multi-modal based on Concatenate and MLSum and present them as our multi-modal baselines, detailed in Section 5.5.2. The results show that multi-modal-based methods significantly outperform unimodal-based methods, illustrating that the model can obtain more task-relevant cues from the multi-modal inputs. In addition, our proposed CPCF achieves remarkable performance gains on different detectors, and our methods surpass our baselines and other state-of-the-art methods by a large margin. For example, our YOLOXCPCF outperforms RGB and Thermal-based YOLOX by 19.3% and 5.7% on mAP_{50}. Also, compared to our multi-modal baselines, the method exceeds the Concatenate and MLSum-based fusion methods by 4.7% and 5.4% on mAP_{50}, 3.6% and 3.4% on mAP which shows the advancement and efficiency of our CPCF. Notably, in our multi-modal baselines, the methods based on MLSum outperform those based on concatenation on nearly all metrics. This further illustrates that compared to directly concatenating inputs from different modalities, using an intermediate fusion strategy is more effective in extracting multi-modal information, thereby enhancing the performance of the model. We also observe the consistent performance boosts of our method on other types of detectors, which reflects the effectiveness of our method as well as its strong generalization capability. In addition, our YOLOv5CPCF and YOLOXCPCF also outperform existing multi-modal methods.

Table 5.3 presents the results of our methods and the competing methods on the LLVIP dataset. As we can see, a good result can be achieved even if only one modality is used. For example, YOLOX using only thermal data achieves a mAP of 60.6. This is due to the relatively simple nature of the dataset scenarios, which does not necessitate differentiating among various categories within the detection targets. Compared to unimodal methods, it is evident that

Table 5.2: Comparison with the state-of-the-art RGB-T fusion methods and our baselines on FLIR dataset by mAP in percentage.

Method	**Backbonne**	**Fusion**	**Modality**	**mAP**$_{50}$	**mAP**$_{75}$	**mAP**	**Param. (M)**
Fcos [62]	ResNet50	-	RGB	59.3	20.2	26.7	32.12
Fcos [62]	ResNet50	-	Thermal	69.4	28.3	33.7	32.12
YOLOv5 [162]	Darknet53	-	RGB	65.2	21.9	29.3	7.03
YOLOv5 [162]	Darknet53	-	Thermal	78.9	32.9	39.2	7.03
YOLOX [63]	Darknet53	-	RGB	62.8	22.2	28.9	8.94
YOLOX [63]	Darknet53	-	Thermal	76.4	36.3	40.2	8.94
Multi-modal methods							
GAFF [157]	ResNet18	GAFF	RGB-T	72.9	32.9	37.5	23.75
CFT [149]	CFB	CFT	RGB-T	78.7	35.5	40.2	206.03
YOLOFusion [147]	Darknet53	CMAFF	RGB-T	76.6	-	39.8	12.52
UA-CMDet [146]	Darknet53	UA-CM	RGB-T	78.6	-	-	-
CSAA [148]	ResNet50	CSAA	RGB-T	79.2	37.4	41.3	-
Our baselines							
FcosCAT	ResNet50	Concatenate	RGB-T	68.0	25.5	32.1	32.13
FcosSUM	ResNet50	MLSum	RGB-T	70.4	28.9	34.5	55.63
YOLOv5CAT	Darknet53	Concatenate	RGB-T	77.0	31.5	38.1	7.03
YOLOv5SUM	Darknet53	MLSum	RGB-T	79.2	34.6	40.2	11.2
YOLOXCAT	Darknet53	Concatenate	RGB-T	77.4	36.9	41.0	8.94
YOLOXSUM	Darknet53	MLSum	RGB-T	76.7	37.7	41.2	13.15
Our implementation with CPCF							
FcosCPCF	ResNet50	CPCF	RGB-T	73.4	32.0	37.0	61.28
YOLOV5CPCF	Darknet53	CPCF	RGB-T	81.6	37.0	41.8	12.67
YOLOXCPCF	Darknet53	CPCF	RGB-T	**82.1**	**41.2**	**44.6**	14.61

Table 5.3: Comparison with the state-of-the-art RGB-T fusion methods and our baselines on LLVIP dataset by mAP in percentage.

Method	**Backbonne**	**Fusion**	**Modality**	**mAP**$_{50}$	**mAP**$_{75}$	**mAP**
Fcos [62]	ResNet50	-	RGB	86.8	45.2	46.5
Fcos [62]	ResNet50	-	Thermal	94.2	62.1	57.4
YOLOv5 [162]	Darknet53	-	RGB	88.0	47.8	48.0
YOLOv5 [162]	Darknet53	-	Thermal	94.7	62.4	58.2
YOLOX [63]	Darknet53	-	RGB	89.3	48.3	48.6
YOLOX [63]	Darknet53	-	Thermal	94.4	67.3	60.6
Multi-modal methods						
ECISNet [158]	ResNet50	ECIS	RGB-T	95.7	-	-
UA-CMDet [146]	Darknet53	UA-CM	RGB-T	96.3	-	-
CSAA [148]	ResNet50	CSAA	RGB-T	94.3	66.6	59.2
CFT [149]	CFB	CFT	RGB-T	**97.5**	72.9	63.6
Our baselines						
FcosCAT	ResNet50	Concatenate	RGB-T	94.5	61.6	57.9
FcosSUM	ResNet50	MLSum	RGB-T	95.1	64.8	58.5
YOLOv5CAT	Darknet53	Concatenate	RGB-T	95.1	62.7	58.2
YOLOv5SUM	Darknet53	MLSum	RGB-T	95.6	65.8	59.4
YOLOXCAT	Darknet53	Concatenate	RGB-T	93.4	65.8	58.1
YOLOXSUM	Darknet53	MLSum	RGB-T	93.4	69.0	61.0
Our implementation with CPCF						
FcosCPCF	ResNet50	CPCF	RGB-T	96.0	69.5	60.6
YOLOv5CPCF	Darknet53	CPCF	RGB-T	96.1	70.1	62.0
YOLOXCPCF	Darknet53	CPCF	RGB-T	96.4	**75.4**	**65.0**

Table 5.4: Comparison with the state-of-the-art RGB-T fusion methods and our baselines on DroneVehicle dataset by mAP in percentage.

Method	**Modality**	**Fusion**	**mAP**$_{50}$	**mAP**$_{75}$	**mAP**
FasterRCNN [52]	RGB	-	63.0	28.6	31.4
FasterRCNN [52]	Thermal	-	71.9	49.6	43.6
RetinaNet [163]	RGB	-	58.0	26.9	29.5
RetinaNet [163]	Thermal	-	66.6	48.2	41.4
S^{2}A-Net [66]	RGB	-	64.1	29.4	32.3
S^{2}A-Net [66]	Thermal	-	74.4	52.5	45.9
PSC [152]	RGB	-	66.9	32.0	33.8
PSC [152]	Thermal	-	75.3	54.8	46.9
Multi-modal methods					
UA-CMDet [146]	RGB-T	UA-CM	63.3	-	-
ECISNet [158]	RGB-T	ECIS	76.0	-	-
Our baselines					
FasterRCNNCAT	RGB-T	Concatenate	74.1	49.5	44.4
FasterRCNNSUM	RGB-T	MLSum	74.7	52.0	45.7
RetinaNetCAT	RGB-T	Concatenate	69.7	49.3	43.1
RetinaNetSUM	RGB-T	MLSum	70.1	51.0	43.8
S^{2}A-NetCAT	RGB-T	Concatenate	75.7	53.2	46.6
S^{2}A-NetSUM	RGB-T	MLSum	76.1	55.8	47.8
PSCCAT	RGB-T	Concatenate	75.6	55.4	47.2
PSCSUM	RGB-T	MLSum	77.3	58.0	48.8
Our implementation with CPCF					
FasterRCNNCPCF	RGB-T	CPCF	76.1	52.8	46.6
RetinaNetCPCF	RGB-T	CPCF	72.9	53.01	45.7
PSCCPCF	RGB-T	CPCF	77.8	**58.1**	49.4
S^{2}A-NetCPCF	RGB-T	CPCF	**79.2**	57.9	**49.7**

multi-modal methods significantly improve the regression accuracy of bounding boxes. For instance, our YOLOX-CPCF shows a marked increase on the mAP$_{75}$ metric, improving by 26.6% over YOLOX (RGB) and 7.6% over YOLOX (Thermal). Moreover, our methods achieve superior performance than our baselines and other existing methods. In addition, we obtain a consistent performance improvement even with other types of detectors.

Different from FLIR and LLVIP datasets, DroneVehicle is a more challenging large-scale dataset targeting oriented object detection in low-light conditions. We compare our methods with the state-of-the-art oriented object detectors on the DroneVehicle dataset and report the experimental results in Table 5.4. Specifically, we modify our detection heads to support detection with orientation for general-purpose object detectors, such as FasterRCNN [52] and RetinaNet [163]. Moreover, we modify their feature extraction architectures for state-of-the-art oriented object detection methods, such as S^{2}A-Net [66] and PSC [152], to accommodate multi-modal inputs. It can be seen that the multi-modal-based methods are considerably better than the unimodal-based methods. For instance, our S^{2}A-NetCPCF demonstrates a significant improvement over methods based on RGB and thermal images, with perfor-

mance increases of 12.3% and 3.9%, respectively. In addition, our multi-modal baselines achieve competitive results compared to existing multi-modal-based state-of-the-art methods, and the model results are further enhanced with the benefit of our proposed CPCF strategy. All the experiments conducted on this dataset validate the versatility of our approach across different types of detectors and its generalizability in various scenarios.

5.5.4.2 Qualitative Results

In Figure 5.5, we compare the detection results of our proposed CPCF with different baselines on the FLIR validation set. In the experiment, we use YOLOX [63] as the base detector to generate single-modality detection results, i.e., RGB-Only and Thermal-Only. Then, for multi-modal fusion, we generated detection results based on Concatenate and MLSum, refer to Section 5.5.2. As shown in the second and third rows of the figure, the RGB images provide rich texture information under clear weather conditions, while thermal images offer more object clues under low-light conditions. It can be seen that the results generated utilizing RGB images are superior to those generated by thermal images under clear weather conditions, which could be attributed to the lack of texture information in thermal images making it difficult to distinguish different individuals in dense objects. This phenomenon is reversed under low-light conditions, illustrating the complementarity between RGB and thermal images. On the other hand, multi-modal methods attempt to leverage this complementarity. As can be seen from the fourth and fifth rows, multi-modal methods clearly outperform unimodal ones. More specifically, a simple concatenation of RGB and thermal images at the input stage can combine information from different modalities to a certain extent, but it falls short when it comes to detecting targets that are unclear in appearance or partially obscured. The use of intermediate fusion strategies can alleviate this issue, but still struggles to handle complex scenarios. CPCF, by employing channel and spatially correlated attention during the intermediate fusion process, effectively utilizes clues from different modalities, achieving the best detection results, as shown in the last row.

Figure 5.6 illustrates the detection results on the DroneVehicle validation set. We use S^2A-Net [66] as our foundational oriented object detection framework. On the other hand, although our multi-modal baseline improves detection results, it still falls short in detecting obscured or densely clustered objects. For instance, the baseline methods lose the obscured vehicle in the first column scenario and fail to identify the densely packed objects in the upper right corner of the scene in the last column. In contrast, our proposed method demonstrates stable results under these scenarios, further attesting to the effectiveness of our approach.

5.5.4.3 Ablation Study

In this section, we conduct ablation experiments on the FLIR dataset for a detailed analysis of our designs. The CPCF consists of three modules: channel-wise cross-attention (CCA), patch-wise cross-attention (PCA), and adaptive gating (AG) module. As presented in Table 5.6, we use YOLOX as a case study and progressively incorporate these

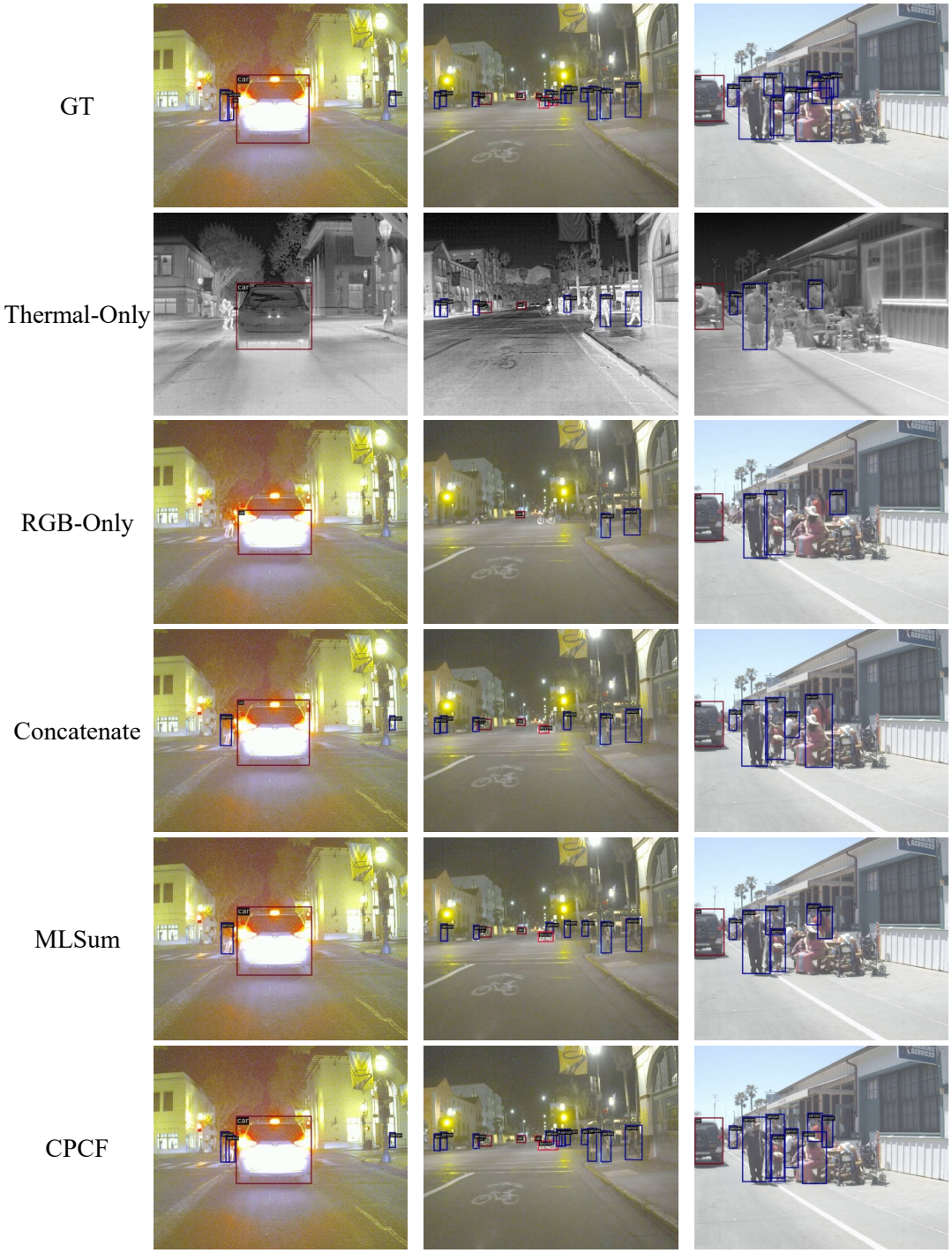


Figure 5.5: Qualitative comparison of four baselines and our proposed method on FLIR validation set. Only bounding boxes with a confidence greater than 0.7 are displayed.

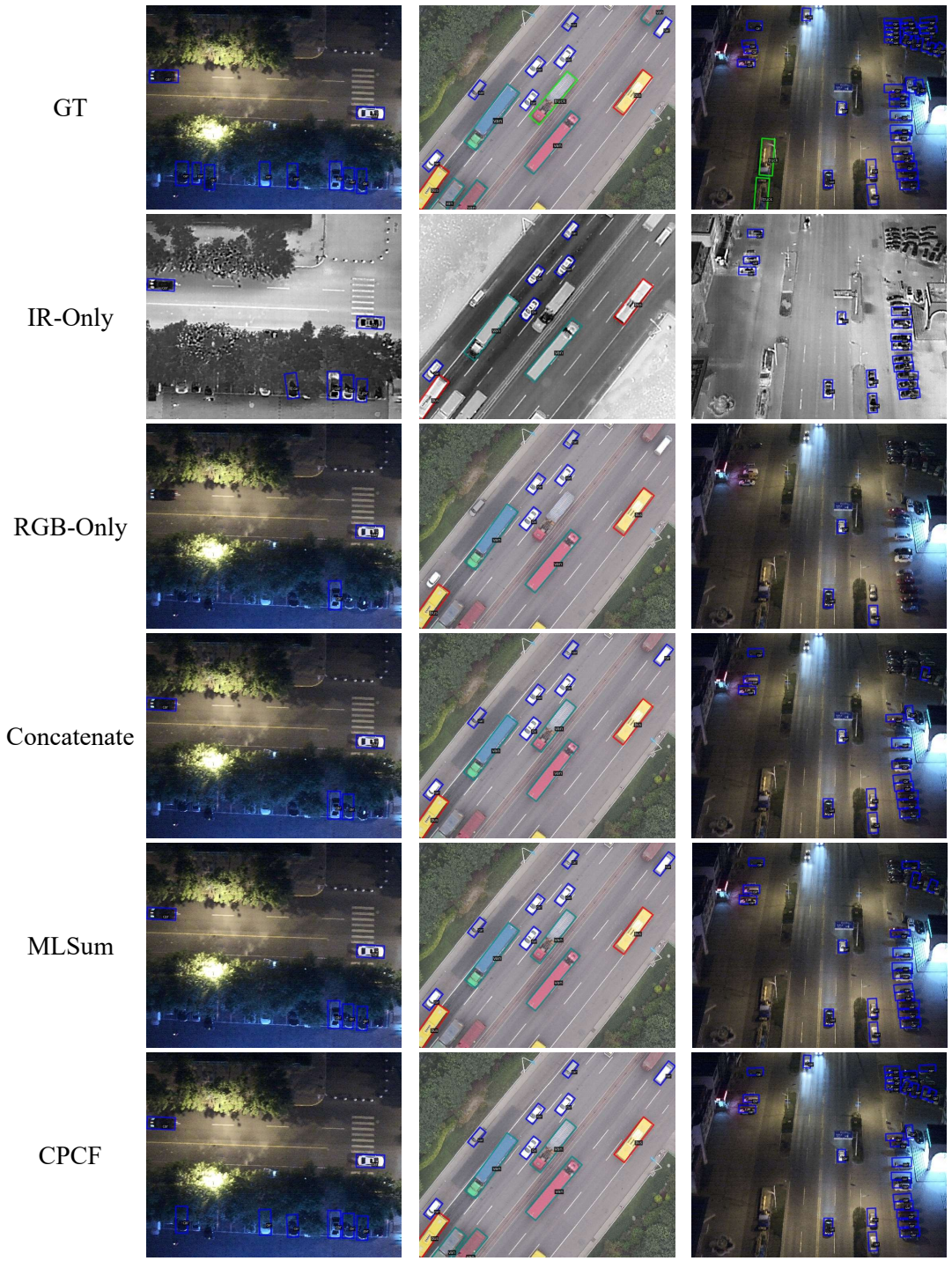


Figure 5.6: Qualitative comparison of four baselines and our proposed method on DroneVehicle validation set. Only bounding boxes with a confidence greater than 0.7 are displayed.

Table 5.5: Comparision of model parameters and flops and fps.

Detector	Modality	Param. (M)↓	FLOPs (G)↓	Runtime (ms)↓
YOLOv5	RGB/T	7.03 (-4.17)	6.35 (-4.17)	20.3 (-3.2)
YOLOv5SUM	RGB-T	11.20 ($\pm$0.0)	10.52 ($\pm$0.0)	23.5 ($\pm$0.0)
YOLOv5CCA	RGB-T	11.26 (+0.06)	10.53 (+0.01)	27.1 (+3.6)
YOLOv5PCA	RGB-T	12.60 (+1.40)	10.60 (+0.08)	29.1 (+5.6)
YOLOv5CPCF	RGB-T	12.66 (+1.46)	10.61 (+0.09)	35.1 (+11.6)
YOLOX	RGB/T	8.94 (-4.21)	10.66 (-4.38)	11.6 (-4.3)
YOLOXSUM	RGB-T	13.15 ($\pm$0.0)	15.04 ($\pm$0.0)	15.9 ($\pm$0.0)
YOLOXCCA	RGB-T	13.22 (+0.07)	15.05 (+0.01)	18.9 (+3.0)
YOLOXPCA	RGB-T	14.55 (+1.40)	15.12 (+0.08)	21.5 (+5.6)
YOLOXCPCF	RGB-T	14.61 (+1.46)	15.13 (+0.09)	26.7 (+10.8)

Table 5.6: Ablation study of the components of our CPCF on FLIR dataset. ● and ○ indicate activated and inactivated components, respectively.

Method	CCA	PCA	AG	mAP_{50}	mAP_{75}	mAP
YOLOXSUM	○	○	○	76.7	37.7	41.2
YOLOXCCA	●	○	○	80.7 (+4.0)	38.6 (+0.9)	43.0 (+1.8)
YOLOXPCA	○	●	○	80.8 (+4.1)	39.8 (+2.1)	43.1 (+1.9)
YOLOXCPCF	●	●	○	81.1 (+4.4)	39.9 (+2.2)	43.4 (+2.2)
YOLOXCPCF	●	●	●	**82.1 (+5.4)**	**41.2 (+3.5)**	**44.6 (+3.4)**

modules into the model to investigate their individual contributions to the overall performance. Specifically, we employ YOLOXSUM as our multi-modal baseline for a fair comparison, as shown in the first row of the table. We then replace the summation operation in YOLOXSUM with CCA and PCA respectively. The results from the second and third rows show that the model's performance in terms of mAP improved by 1.8% and 1.9% with the application of CCA and PCA, respectively. Finally, to illustrate the role of AG, we conduct experiments using fixed weights of 0.5 and dynamic weights with AG and obtain performance boosts of 2.2% and 3.4%, respectively, as shown in the last two rows of the table. The results reveal that compared to manually setting fixed weights, employing AG can greatly enhance the model's performance. This further suggests that different weights should be assigned to different attention mechanisms at various stages of the model to adapt to the changes in information volume in the channel and spatial dimensions. Therefore, we conclude that the introduction of CCA and PCA can provide more efficient feature extraction capabilities for intermediate fusion from both channel and spatial dimensions, thereby enhancing model performance. Moreover, the dynamic weight allocation mechanism of AG can further optimize fusion efficiency according to changes of information in channel and spatial dimensions, thereby dealing with complex multi-modal data more effectively.

Table 5.7: Comparison of MLP-based cross-attention and our self-attetnion-based cross-attention on FILR, LLVIP, and DroneVehicle datasets.

Method	Dataset	Cross Attention	mAP$_{50}$	mAP$_{75}$	mAP	Param. (M)
YOLOXCPCF	FLIR	MLP-Based	79.7	38.9	42.3	6.50
		Ours	**82.1**	**41.2**	**44.6**	**1.03**
YOLOXCPCF	LLVIP	MLP-Based	94.8	71.7	62.8	6.50
		Ours	**96.4**	**75.4**	**65.0**	**1.03**
S^{2}A-NetCPCF	DroneVehicle	MLP-Based	78.0	57.1	48.8	6.50
		Ours	**79.2**	**57.9**	**49.7**	**1.03**

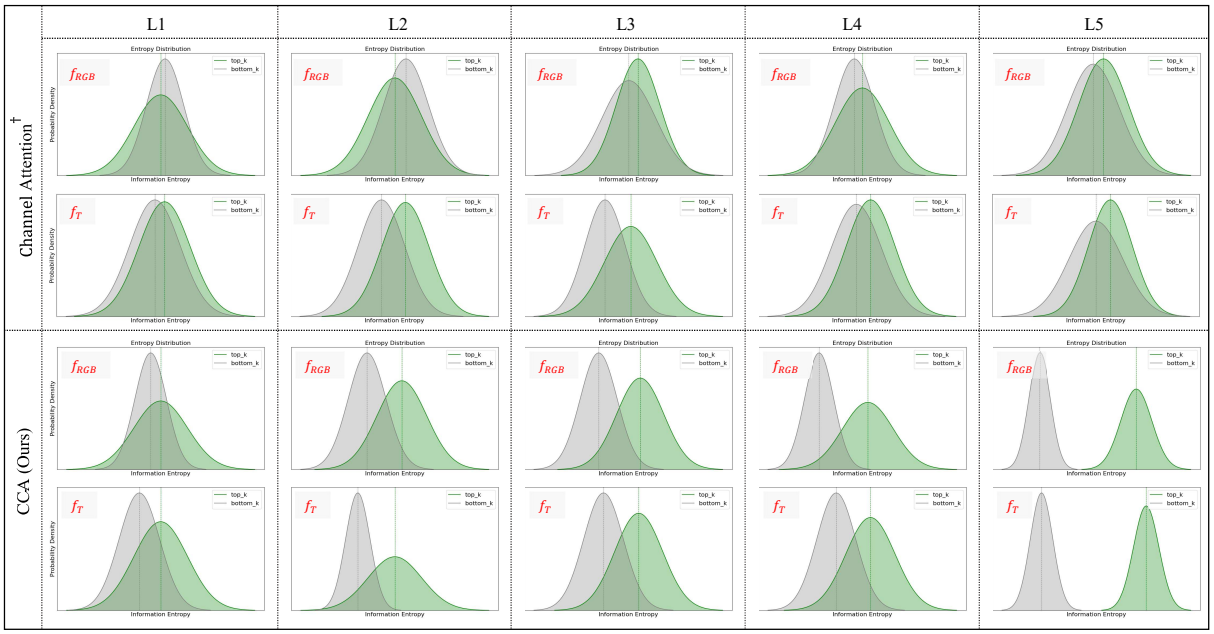


Figure 5.7: Comparison of information entropy distributions of top and bottom 16 channels of RGB and Thermal feature maps at different levels. † denotes MLP-based cross-attention.

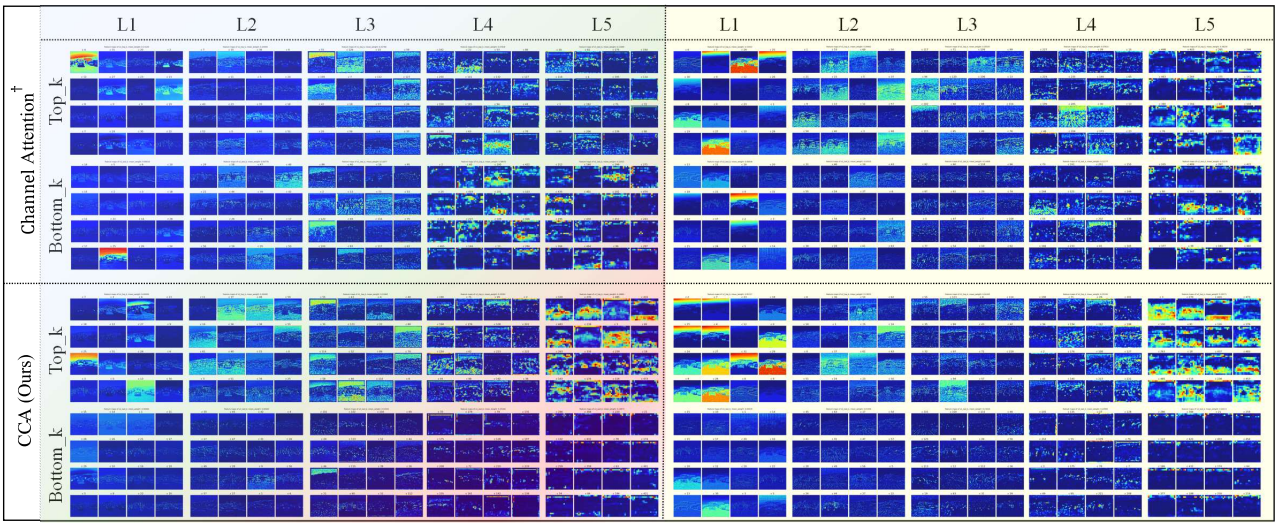


Figure 5.8: Visualization of top_k and bottom_k channel features.The left side shows the RGB data stream at different stages, while the right side shows the counterpart of the thermal data stream. † denotes MLP-based cross-attention.

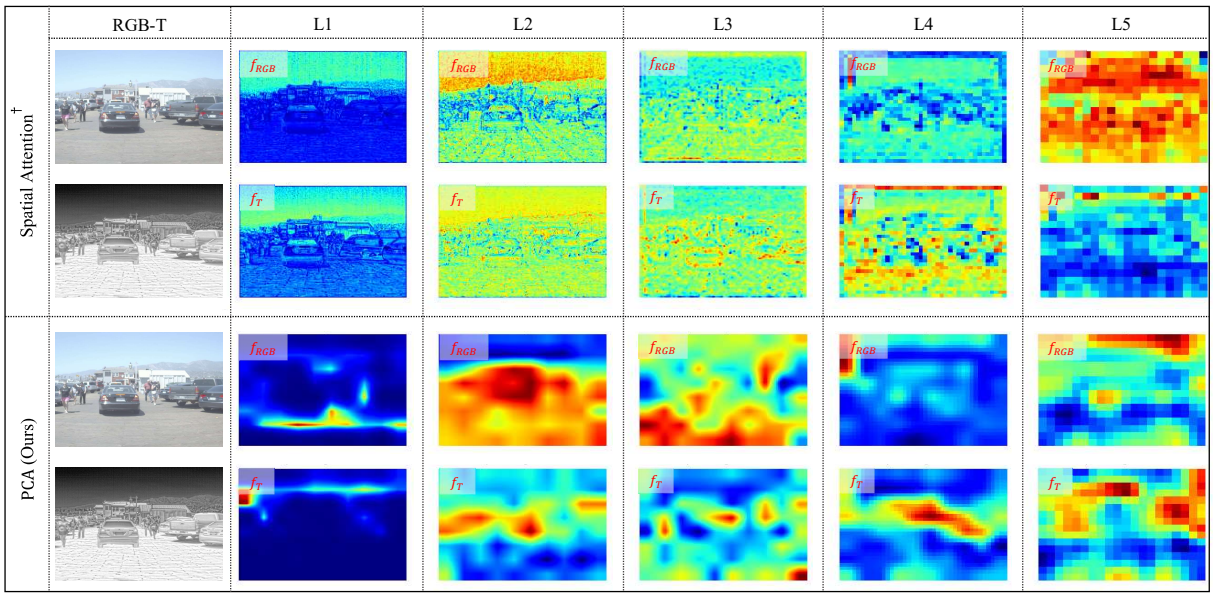


Figure 5.9: Comparison of spatial attentions of RGB and Thermal feature maps at different levels. † denotes MLP-based cross-attention.

5.5.5 Attention Analysis

To further illustrate the effectiveness of the proposed cross-attention mechanism, we employ MLP-based channel and spatial attention to replace our CCA and PCA modules. Notably, the MLP, which projects concatenated feature maps into 1-D attention maps [106], is widely used in various attention mechanisms, such as [147, 158]. As shown in Table 5.7, our proposed self-attention-based CCA and PCA significantly outperform the MLP-based attention mechanisms on different datasets. Moreover, we also quantify the parameters of a single standalone cross-attention module, which takes a feature map of size 128×168 with 512 channels as input. As shown in the last column of the table, compared to the MLP-based cross-attention, our method saves approximately 85% of the parameters, proving its higher efficiency.

In Figure 5.7, we compare the information entropy of different feature channels at different stages depicted in Figure 5.1. Specifically, we first rank the channels in the feature map according to the channel attention scores, and then calculate the information entropy of the top 16 feature channels and the bottom 16 feature channels, termed top_k and bottom_k, respectively. In the figure, the green distribution describes the information entropy of top_k, while the gray corresponds to bottom_k. We observe that, in the MLP-based channel attention, the information entropy distribution of top_k and bottom_k is strikingly similar. This suggests that the amount of information in top_k and bottom_k is consistent. Therefore, when the attention score of bottom_k is very low, many channels containing valuable information might be suppressed. Our method, on the other hand, concentrates valuable information into a subset of channels and better recognizes these channels with the CCA module, thereby suppressing redundant information while allocating more attention to channels with more information. Besides, it is evident that top_k

always contains more information. Particularly in the final stage, as illustrated in the column of L5, the discrepancy between the information distribution of top_k and bottom_k in high-level semantic feature maps is further amplified, demonstrating the effectiveness of our CCA in channel awareness. More details can be seen in Figure 5.8.

Additionally, Figure 5.9 demonstrates the spatial attention maps in different cross-attention mechanisms. It is noticeable that, compared to MLP-based spatial attention, our PCA pays more attention to the areas of interest, and the attentions of different modalities complement each other to a certain extent, which indicates that our method is able to utilize the complementarity between modalities more efficiently while forming spatial awareness.

5.5.6 Speed and Parameter Analysis

To further assess the practicality of our proposed fusion method, we choose the widely used single-stage detectors YOLOv5 and YOLOX as benchmarks and conduct tests to measure the execution speed of our method. In Table 5.5, we report the total number of learnable parameters, the number of floating-point operations (FLOPs), and the runtime. All models in the experiments are implemented based on MMDetection [161], and running on a laptop equipped with an RTX2080 GPU. It can be observed that multi-modal methods show a decrease in speed compared to unimodal methods. For instance, the runtime of YOLOv5SUM and YOLOXSUM increased by approximately 3ms compared to the single-modality counterparts. This is due to the fact that intermediate fusion introduces additional feature extraction branches, leading to an increase in computational complexity. Additionally, the use of our lightweight fusion module results in a minor increase in runtime. Specifically, CCA adds approximately 3ms, while PCA contributes an extra 5ms. When combining both CCA and PCA, i.e., our CPCF, the runtime increases by around 10ms. Furthermore, compared to our multi-modal baseline, our fusion strategy adds virtually no extra parameters or floating-point operations. In addition, in the last row of Table 5.2, we list the number of parameters for different models. It is evident that our method manages to achieve state-of-the-art performance while ensuring the model remains lightweight.

5.6 Summary

In this Chapter, we present a lightweight multi-modal cross-fusion method termed CPCF for visible-infrared object detection, which consists of channel-wise cross-attention (CCA), patch-wise cross-attention (PCA), and an adaptive gating (GA) module. The CCA and PCA are designed to refine valuable cues from the channel and spatial dimensions, respectively, and operate the features of one modality to calibrate another, thereby better integrating the information of different modalities. Moreover, we argue that the useful multi-modal information contained within channel and spatial dimensions can vary during the forward propagation process. To account for this, we design the AG module to adaptively adjust the attention weights in the channel and spatial dimensions. Subsequently, based on the

CPCF, we design a universal intermediate fusion architecture that allows for extension to various types of detectors, facilitating the harnessing of multi-modal information to enhance the model's performance. Finally, we conduct extensive experiments with various object detection frameworks on standard and oriented object detection datasets. The results demonstrate that our method is able to effectively capture information from different modalities and consistently outperform other advanced multi-modal methods. Additionally, thanks to its lightweight design, our method can be incorporated into lightweight object detection models, enabling real-time object detection.

Chapter 6

Multi-modal Unsupervised Domain Adaptation

In the preceding chapters, we have thoroughly explored the performance of deep perceptual models based on multi-modal visual fusion in the context of semantic segmentation and object detection tasks. However, in real-world applications, it is often challenging to fully exploit the available data, especially the unlabeled ones, due to the lack of sufficient annotated data to support fully supervised learning. Hence, in this chapter, we shift our focus toward a more challenging scenario, where there is no labeled data on the target dataset, and explore how to train a model on it. In Chapter 2, we conducted a review of unsupervised learning methods based on domain adaptation techniques and observed that such practices could effectively leverage the supervisory signals provided by the related domains to establish valuable decision priors for the target domain. In this context, this chapter mainly investigates how to leverage multi-modal information to facilitate the model in learning domain-independent feature representations in the domain adaptation process, and to increase the inter-classes distance in the semantic feature space while closing the gap between domains. In addition, given that a simulator can generate an arbitrary amount of accurately labeled data, this chapter regards the synthetic dataset with its precise labels as the source domain. Conversely, the real-world dataset without labels is considered the target domain.

6.1 Abstract

We propose a novel multi-modal-based Unsupervised Domain Adaptation (UDA) method for semantic segmentation. Recently, depth has proven to be a relevent property for providing geometric cues to enhance the RGB representation. However, existing UDA methods solely process RGB images or additionally cultivate depth-awareness with an auxiliary depth estimation task. We argue that geometric cues that are crucial to semantic segmentation, such as

local shape and relative position, are challenging to recover from an auxiliary depth estimation task with mere color (RGB) information. In this paper, we propose a novel multi-modal UDA method named MMADT, which relies on both RGB and depth images as input. In particular, we design a Depth Fusion Block (DFB) to recalibrate depth information and leverage Depth Adversarial Training (DAT) to bridge the depth discrepancy between the source and target domain. Besides, we propose a self-supervised multi-modal depth estimation assistant network named Geo-Assistant (GA) to align the feature space of RGB and depth and shape the sensitivity of our MMADT to depth information. We experimentally observed significant performance improvement in multiple synthetic to real adaptation benchmarks, i.e., SYNTHIA-to-Cityscapes, GTA5-to-Cityscapes and SELMA-to-Cityscapes. Additionally, our multi-modal UDA scheme is easy to port to other UDA methods with a consistent performance boost.

6.2 Introduction

Semantic segmentation is a fundamental task in computer vision which aims to assign a label to each pixel in an image. The past few years have witnessed a significant progress of fully supervised approaches because of the advent of advanced deep networks, e.g.,[164] and large-scale, well-annotated datasets, e.g., [31]. However, the performance of these methods in real world scenarios highly depends on the similarity of the test scene to the training images, which is improper in practical applications. Besides, collecting enough images with fine-grained annotations is hugely labor-intensive, e.g., building a single densely annotated image in the Cityscapes dataset [31] takes around 1.5 hours. The common practice to circumvent such problems is to fit a model on readily-accessible synthetic data with annotations, then adapt it to the target-specific data, referred to as domain adaptation.

Unsupervised domain adaptation (UDA) strives to optimize a model under the supervision of the source domain while obtaining the lowest prediction error on the target domain, in which no annotated data is available. In recent years, thanks to adversarial training and self-training strategies, UDA has made considerable progress. Inspired by multi-task learning (MTL) [165], several works have turned to consider the use of auxiliary tasks to assist adaptation. Typically, the depth image is employed as privileged information to train an auxiliary task of depth estimation [166], as depth information is easy-to-access and tightly coupled with semantic information. Consequently, leveraging depth information is shown to be an effective way to address the UDA challenge. Yet, previous depth aware methods define depth under the MTL paradigm, i.e., learning and inferring depth hints from color images with the usage of a depth estimation auxiliary task. We argue that depth clues that complement colors are hard to deduce from color images alone, thus existing methods fail to capture valid geometric information.

Multi-modal learning (MML) refers to combining the related information of different modalities to improve the model's representation. For multi-modal semantic segmentation, RGB-D data are commonly used to explore complementary cues of color and geometry and has achieved superior results than RGB-only data . Thus, it is of practical value to investigate the use of additional modalities to lift the UDA method under the MML paradigm. However, exist-

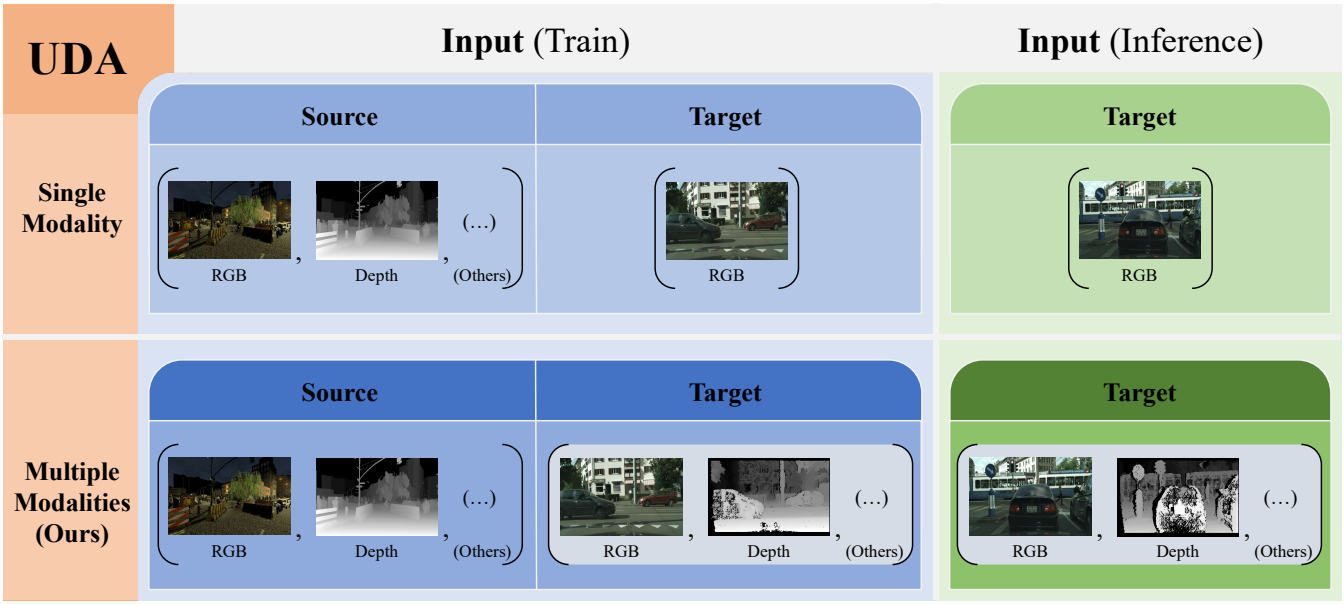


Figure 6.1: Multi-Modal UDA input data. The upper part shows general UDA training and testing input. The lower part shows our multi-modal UDA training and testing input.

ing UDA-based semantic segmentation models process RGB images solely and adaptations under the MML paradigm are still underestimated. To this end, this work focuses on improving the adaptive capability with multi-modal visual data. Practically, our UDA scheme relies on both RGB and depth images as input, as illustrated in Figure 6.1. We view the depth image from the source and target domain as an additional modality and explore valuable geometric cues directly from the raw depth to guide and constrain the adaptation process. In addition, as the raw depth of the target domain is captured with considerable noise and has a different depth range from the source domain, introducing depth input will carry a vast discrepancy in depth images of different domains, as shown in Figure 6.1, which further exacerbates domain shift. To address this problem, we propose to bridge such discrepancy in three ways: 1) We design an efficient Depth Fusion Block (DFB) to recalibrate the input depth and align it with the RGB features. 2) We explicitly align the feature distribution of depth by Depth Adversarial Training (DAT). 3) We present a self-supervised multi-modal depth estimation assistant network named Geo-Assistant (GA) to transfer the geometric attention to our UDA model. Concretely, we design a data augmentation strategy named Complementary Random Cutout (CRC) for training GA, which is then used to align RGB and depth features and improve the sensitivity of our UDA models to depth information. Equipped with these designs, we propose a novel multi-modal UDA training scheme and a multi-modal UDA model named MMADT. To the best of our knowledge, this is the first work to focus on multi-modal UDA for semantic segmentation tasks.

In addition, with its elegant design, our MMADT is capable of being well aware of the geometric clues and binding them to the RGB features, thus decreasing the domain shift existing in both modalities. As a result, our method significantly improves adaptation performance and performs favorably against RGB-only based methods on SYNTHIA-

to-Cityscapes, GTA5-to-Cityscapes, and SELMA-to-Cityscapes UDA settings. Our contributions can be summarized as follows:

- We offer a multi-modal UDA method for semantic segmentation and train a multi-modal UDA model (MMADT), which takes both RGB and depth images as input to improve the performance of the UDA network.
- We propose Depth Fusion Block (DFB) and Depth Adversarial Training (DAT) to bridge the discrepancy of depth between the source and target domain while fusing depth cues into RGB features.
- We propose a Complementary Random Cutout (CRC) data augmentation method tailored for self-supervised multi-modal Geo-Assistant (GA), which can help align features and learn complementary information from different modalities.

This paper is organized as follows. Section 6.3 reviews the existing works related to our method. The overall framework is presented in section 6.4 with the details of DFB, DAT and GA. The experimental set up and the results are presented and discussed in section 6.5. Finally, Section 6.7 ends this paper with conclusion and discussion.

6.3 Related Works

6.3.1 UDA-based semantic segmentation

Unsupervised semantic segmentation can be seen as a branch of the UDA problem, which aims at pixel-wise classification in the UDA setting. Generally, UDA-based segmentation methods can be categorized into adversarial training and self-training. Adversarial training aims to align the domain distribution by optimizing a discriminator to grow the domain confusion, which can be performed in several ways. Hoffman et al. [167] studied to align domains at the feature level, which is the first work to pay attention to urban scene UDA semantic segmentation. Then, Vu et al. [168] investigated boosting the performance at the output or patch level. Rui et al. [169] proposed to transfer the domain appearance style from the source to the target via image translation to increase domain confusion. Dayan et al. [170] introduced multi-scale consistency regularization to improve the stability of model predictions. On the other hand, self-training seeks to optimize the UDA model by considering pseudo-labels of the target domain. As a pioneer work, Zou et al. [171] recursively refined the pseudo-labels by solving class imbalance and introducing spatial priors. Then, Zheng et al. [172] picked high confidence predictions to improve the accuracy of pseudo-labels, while Zou et al. [173] further applied model regularization to mitigate the impact induced by incorrect pseudo-labels. Moreover, Wilhelm et al. [174] mixed images from two domains according to their corresponding labels and pseudo-labels to stabilize the training procedure.

Due to the high correlation between geometric and semantic information, some recent work advocates implicitly learning geometry-related information through auxiliary tasks. Lee et al. [166] proposed a method named SPIGAN

that treated depth as the privileged information and defines an auxiliary task for depth estimation, which was jointly optimized with the segmentation task. Moreover, Chen et al. [100] considered depth estimation as an auxiliary task while training the style transfer network with the help of depth information at the input level. The method named DADA [19] further integrated the predicted depth into the segmentation network at the output level and leveraged adversarial training to bridge the performance gap. Wang et al. [18] proposed CorDA which considers depth information of both source and target domain, in which the depth of target domain was prepared off-line by an off-the-shelf approach. Unlike existing depth-aware methods, our MMADT takes both RGB and raw depth images as input in both the training and testing phases and is optimized in a multi-modal learning manner rather than a multi-task learning manner.

6.3.2 RGB-D semantic segmentation

RGB-D semantic segmentation refers to the use of RGB and depth images as multi-modal input and optimize a semantic segmentation task by leveraging the privileged information of different modalities. Basically, according to the stages of multi-modal information fusion, the RGB-D semantic segmentation methods can be divided into three groups: early fusion, intermediate fusion, and late fusion methods [14]. In particular, the early fusion practices try to combine RGB and depth information at the input level. For example, Xing et al. [175] took the concatenated RGB-D as an input and re-customize the convolution operation to efficiently extract semantically relevant geometric clues from the input. The intermediate fusion strategies tend to encode geometric features through a depth feature extractor and integrate multi-modal information at the feature level. Seichter et al. [176] fused the features at multiple encoding stages and utilizes factorized convolutions to create a lightweight model for real-time operations. Zhou et al. [177] proposed to find better interaction patterns for RGB and depth information through customized attention mechanisms. Finally, late fusion approaches feed RGB and depth into two parallel networks and merge the information at the output level. Valada et al. [76] proposed deep expert fusion techniques to fuse multiple decision information in the output stage, while Cheng et al. [73] leveraged gated fusion layer to estimate the weight of multiple experts on decisions.

Empirically, fusing features at the intermediate level allows both flexible control of the fusion process and computational reduction of forward-pass [14]. Accordingly, our method combines the information of different modalities at the feature level.

6.3.3 Self-supervised learning

Self-supervised learning (SSL) aims to learn a generalized and semantically meaningful representation with the guidance of the data itself. An intuition behind SSL is to artificially define a pretext task and supervise the model by the pretext-related ground truth. In recent years, a wide range of pretext tasks have been investigated, such as image

inpainting [178]. Another intuition is to learn consistent representations from distorted input [179]. For example, contrastive learning approach [180] optimized a useful embedding space in which similar sample pairs stay close while dissimilar ones are far apart. Further, Zhu et al. [181] introduced contrastive learning into the episode training process to learn category-independent discriminative patterns and stable feature representations. And recently, Liu et al. [182] proposed to learn better transferable visual representation by regarding image rotation prediction as an auxiliary task. In this work, we consider the depth estimation as a pretext task and design a multi-modal data augmentation method to encourage our model to explicitly capture intricate dependencies of RGB-D data.

6.3.4 Knowledge transfer

Knowledge transfer refers to passing knowledge from a complex teacher pretrained on the same or a similar task to a simpler student. Hinton et al. [183] proposed to learn the consistency of output distributions through soft-target. Heo et al. [184] tried to construct a deeper expert and then distilled knowledge to a smaller student by mimicking the intermediate features. Zhang et al. [185] allowed an ensemble of students to learn simultaneously and teach each other throughout the learning procedure. Motivated by attention transfer, we propose a multi-modal network named Geo-Assistant (GA) to efficiently align information of RGB and depth images and integrate color and geometric features. In addition, the GA uses the same encoder architecture as the student network to facilitate the transfer of geometric attention.

6.4 Methodology

In this work, we aim to train a multi-modal semantic segmentation model by leveraging diverse clues of RGB and depth under the UDA setup. To this end, we define the synthetic data as the source domain $\mathcal{X}^S \in S$ and the real-world data as the target domain $\mathcal{X}^T \in T$, where the source and target domain share the same label space. In the following, Section 6.4.1 illustrates the overview of the proposed multi-modal UDA scheme and details the Depth Fusion Block (DFB) and Depth Adversarial Training (DAT) components. Then, we present the Complementary Random Cutout (CRC) data augmentation method and self-supervised Geo-Assistant (GA) in Section 6.4.2. In Section 6.4.3, we combine all components and describe our MMADT training protocol.

6.4.1 Multi-modal UDA Overview

Previous UDA methods take a single modality as input, i.e., RGB image, or consider auxiliary modality, such as depth, only at the training stage. In multi-modal UDA setup, we consider depth as an extra modality in both source and target domains, which is always available during training and testing. Formally, given the source dataset $\mathcal{X}^S = \{x^S_{rgb}, x^S_d\}_{i=1}^{N_S}$ and one-hot labels $\mathcal{Y}^S = \{y^S\}_{i=1}^{N_S}$, we aim to optimize a function to minimize the prediction error on

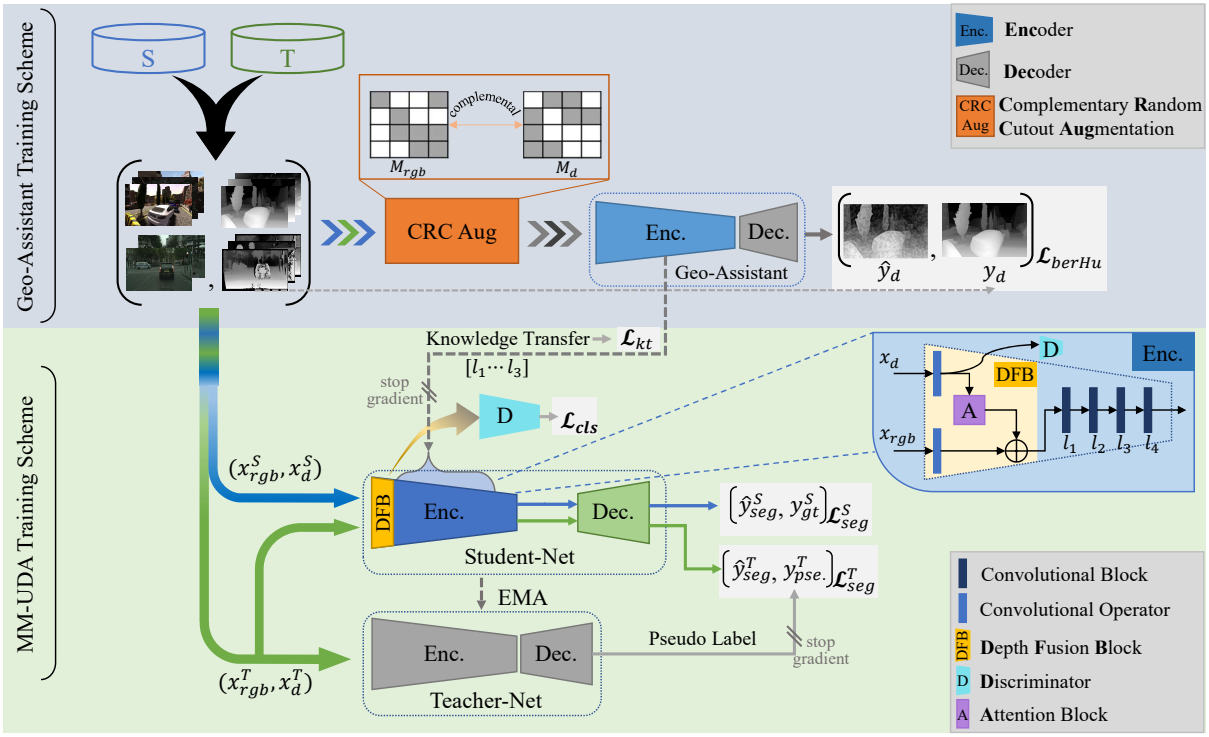


Figure 6.2: Overview of multi-modal UDA scheme. In the top part, we illustrate the self-supervised Geo-Assistant training scheme. In the lower part, we detail the UDA training scheme. Note that both of them take multi-modal, i.e., RGB and depth, as input.

the target dataset $\mathcal{X}^T = \{x^T_{rgb}, x^T_d\}_{i=1}^{N_T}$ where ground truth labels $\mathcal{Y}^T$ are not accessible. To do so, we propose a novel multi-modal UDA training scheme that can benefit from three designs, namely DFB, DAT, and GA, as shown in the lower part in Figure 6.2. Note that our UDA scheme takes both RGB and depth as input.

6.4.1.1 Self-training for UDA

Given a semantic segmentation network U parameterized by θ_U, our UDA training scheme follows the self-training protocol. Naively, we first optimize U with cross-entropy loss on the source domain:

$$\min_{\theta_U} \mathcal{L}^S_{seg}(U) = \mathbb{E}_{(x^S_{rgb}, x^S_d, y^S) \sim (\mathcal{X}^S, \mathcal{Y}^S)} \mathcal{H}[U(x^S_{rgb}, x^S_d | \theta_U), y^S], \tag{6.1}$$

where $\mathcal{H}(\cdot,\cdot)$ refers to cross entropy loss, $U(\cdot,\cdot) \in \mathbb{R}^{(C,H,W)}$ indicates the output of network U, C is the number of semantic classes, H and W are the height and width of input images. Then, in order to adapt the knowledge from the source to the target domain, a network, called Teacher-Net, parameterized by θ_{MT} is defined to generate pseudo-labels $\mathcal{Y}^T \in \mathbb{R}^{(C,H,W)}$ of the target domain:

$$\begin{aligned}\tilde{y}^{T(c,i,j)} &= [c = \underset{c'}{\operatorname{argmax}}\, U(x_{rgb}^T, x_d^T|\theta_{MT})^{(c',i,j)}],\\ \min_{\theta_U} \mathcal{L}_{seg}^T(U) &= \mathbb{E}_{(x_{rgb}^T, x_d^T)\sim\mathcal{X}^T}\mathcal{H}[U(x_{rgb}^T, x_d^T|\theta_U), \tilde{y}^T],\end{aligned} \tag{6.2}$$

where the label of each pixel is a discrete one-hot vector and the Teacher-Net is a twin of network U. Thus, the output of Teacher-Net can be described as $U(x_{rgb}^T, x_d^T|\theta_{MT})$. At training step t, the parameters of Teacher-Net (θ_{MT}) are successively updated by the exponential moving average (EMA) [186] of θ_U:

$$\theta_{MT_t} = \eta\theta_{MT_{t-1}} + (1-\eta)\theta_{U_{t-1}}, \tag{6.3}$$

where η is a hyper-parameter to smooth the update process, which we set to $\eta = 0.999$. In addition, we adopt ClassMix [187], color jitter, and Gaussian blur as data augmentation during the self-training phase as in [174, 18] to stabilize the training procedure.

6.4.1.2 Depth Fusion Block

To elicit the geometric clues, DFB takes depth as input. As illustrated in Figure 6.2, DFB consists of a depth stream and a RGB stream. For the depth stream, the depth input x_d is first encoded by the convolutional operators. Then, the extracted depth features are fed into an attention block to disentangle semantic-meaningful information from noisy features in a process which is called calibration. Following, the calibrated depth features are merged with features from the RGB branch by an element-wise sum. The RGB stream adopts a symmetrical structure with the depth stream, without the attention block. The whole process can be formulated as:

$$\begin{aligned}F_d &= f_s(x_d|\theta_\phi),\\ F_{rgb} &= f_s(x_{rgb}|\theta_\psi),\\ F_{fuse} &= F_{rgb} \oplus F_d,\end{aligned} \tag{6.4}$$

where $f_s(\cdot)$ denotes a sub-network containing part modules of the network U, $\{\theta_\phi, \theta_\psi\} \in \theta_U$ stand for the parameters of the depth stream and RGB stream, respectively. The fusion process of RGB and depth is simple yet efficient, thus avoiding the notable increase in the running burden of the network due to the additional depth branch.

6.4.1.3 Depth Adversarial Training Scheme

DAT is designed to align the distributions of x_d^S and x_d^T. To do so, a discriminator D and an adversarial loss $\mathcal{L}_{adv}$ are defined. Formally, given a depth input x_d, we feed-forward the output of the depth stream (parameterized by θ_ϕ) as input and train the discriminator to make x_d^S and x_d^T distinguishable (source *vs.* target).

$$\begin{aligned}\min_{\theta_D} \mathcal{L}_{cls}(D) = \mathbb{E}_{x_d^S \sim \mathcal{X}^S}[logD(f_s(x_d^S|\theta_{\phi_a}))] \\ + \mathbb{E}_{x_d^T \sim \mathcal{X}^T}[log(1 - D(f_s(x_d^T|\theta_{\phi_a})))],\end{aligned} \tag{6.5}$$

where $\theta_{\phi_a} \in \theta_\phi$ is the parameters before the attention block in the depth stream. Then θ_{ϕ_a} is updated following the adversarial optimization process to fool the discriminator, which can be summarized as a min-max criterion:

$$\min_{\theta_{\phi_a}} \max_{\theta_D} \mathcal{L}_{adv}(\mathcal{X}^S, \mathcal{X}^T, D, f_s) = \mathbb{E}_{x_d^T \sim \mathcal{X}^T}[log(D(f_s(x_d^T|\theta_{\phi_a})))]. \tag{6.6}$$

Once the depth features are aligned, the discriminator would no longer be able to distinguish whether the input depth image comes form the source or the target domain. In this way, the distributions of depth features yielded by the depth stream from the source and target domains are brought closer together. In addition, DAT can be seamlessly integrated into the self-training process to enable end-to-end training of the entire network.

6.4.2 Self-Supervised Multi-Modal Geo-Assistant

For self-supervised learning, we consider depth estimation as a proxy task. Ideally, we expect a network to be able to perceive geometric cues from input data and be more concerned with relative relationships of objects than absolute depth value. Thus, instead of estimating the real depth value, we normalize the depth value into the range of [0, 1], which also attenuates the depth bias of different domains.

6.4.2.1 Complementary Random Cutout (CRC)

The main motivation of CRC comes from the fact that color and depth information are complementary, e.g., the geometric cues of depth information can be seen as additional hints to extract semantic boundaries from the fine-grained color information. Thereby, we propose the CRC data augmentation strategy for training self-supervised multi-modal GA. Specifically, CRC first divides RGB and depth images into regular non-overlapping patches. Then, we define two complementary masks, M_{rgb} and M_d, to ensure only one modality can be seen for each patch during training, as shown in the CRC part of Figure 6.2. Note that the mask areas are randomly selected along the patch, and the sampling strategy is straightforward: we perform non-repeated sampling according to a uniform distribution. To train a more generalized GA, we use a stronger random regularization to increase the diversity of the samples. They are i) randomly choosing the masking ratio from 0.4 to 1.0 on the fly. ii) setting 40% probability and 20% probability of using unmarked RGB and depth, respectively.

Overall, the advantages brought by CRC is threefold i) depth inputs from different domains are unified into the same output space; ii) features from different modalities can be aligned at multiple levels; iii) network is more sen-

sitive to geometric cues.

6.4.2.2 Geo-Assistant Training

As shown in the top part of Figure 6.2, GA takes RGB and depth images as input and is supervised by normalized depth values. The self-supervised GA reconstructs the depth map by predicting relative depth values for each masked and unmasked patch. We optimize the depth estimation by the reverse Huber loss [19]:

$$\mathcal{L}_{berHu}(\hat{y}, y) = \begin{cases} |y - \hat{y}|, & \text{if } |y - \hat{y}| \leq \tau, \\ \frac{(y-\hat{y})^2+\tau^2}{2\tau} & \text{otherwise}, \end{cases} \tag{6.7}$$

where τ refers to a threshold and is set to $\frac{1}{5}$ of the maximum depth residual by default, y and $\hat{y}$ denote the ground truth depth value and the estimated depth value, respectively. We assign different weights to the data from the source and target domains, thus the objective function can be formulated as follows:

$$\min_{\theta_{GA}} \mathcal{L}_{reg}(GA) = \alpha_1 \cdot \mathcal{L}_{berHu}(\hat{y_d}^S, x_d^S) + \alpha_2 \cdot \mathcal{L}_{berHu}(\hat{y_d}^T, x_d^T), \tag{6.8}$$

where α_1 and α_2 balance the weight of loss in different domains, θ_{GA} denotes the parameters of GA, $\hat{y_d}$ and x_d refer to the predicted depth value and raw normalized depth value, respectively. Note that during the training, the GA learns to infer the depth information of masked patches from the corresponding color hints, and remove the depth bias of unmasked patches from the raw depth while retaining the geometric information. Consequently, introducing raw depth as a geometric prior contributes to reducing learning difficulty and allowing the model to focus more on the mapping between different modalities, refer to Section 6.6.1 for more discussions. The detailed training procedure of self-supervised multi-modal GA is summarized in Algorithm 1.

6.4.3 MMADT Training Protocol

Given the multi-modal UDA network MMADT and the pretrained GA, we expect that the MMADT learns to mimic GA, and thus is more sensitive to depth signals. To this end, we distill the geometric cues from GA to our multi-modal UDA model. In addition, we observe that the attention maps can be seen as an abstract representation of the critical clues w.r.t the specific input. Thus, we regard GA as the teacher and MMADT as the student and transfer knowledge through activation-based attention distillation. Concretely, the attention maps are derived from the activated tensor of each encoding stage during training. We train the student to have similar geometric-aware behavior to the teacher network by minimizing the attention distance. Note that we use the same encoder structure for feature extraction to simplify the transfer procedure. Thus, the optimization objective can be defined as:

Algorithm 1: Self-Supervised Multi-Modal Geo-Assistant Scheme

input : Max iterations N, batch size B, hyper-parameters α_1 and α_2;
$\mathcal{X}^S$: source dataset;
$\mathcal{X}^T$: target dataset;
GA: initialized Geo-Assistant parameterized by θ_{GA}
output: θ_{GA}

training:
for $iter \leftarrow 1$ **to** N **do**
  `// multi-modal sampling`
  Randomly sample $\{x^S_{rgb}, x^S_d\}^B_{i=1}$ from $\mathcal{X}^S$;
  Randomly sample $\{x^T_{rgb}, x^T_d\}^B_{i=1}$ from $\mathcal{X}^T$;
  for *each mini-batch* **do**
    Generate complimentary masks from RGB-D input pairs by CRC data augmentation:
      $(Mask_{rgb}, Mask_d)^{S(T)} \leftarrow \texttt{CRC}((x_{rgb}, x_d)^{S(T)})$
    Generate training data by applying $Mask$ to image pair:
      $(x_{rgb}, x_d)^{S(T)}_{in} = (x_{rgb} \otimes Mask_{rgb}, x_d \otimes Mask_d)^{S(T)}$
    Estimate relative depth value:
      $\hat{y_d}^{S(T)} \leftarrow GA((x_{rgb}, x_d)^{S(T)}_{in})$
    Collect gradients of GA:
      $\nabla^{ga} \leftarrow \texttt{BackProp}(\alpha_1 \cdot \mathcal{L}_{berHu}(\hat{y_d}^S, x^S_d) + \alpha_2 \cdot \mathcal{L}_{berHu}(\hat{y_d}^T, x^T_d))$
    Update θ_{GA} of GA with Adam [188]:
      $\theta_{GA} \leftarrow \texttt{Adam}(\nabla^{ga}, \theta_{GA})$

$$
\begin{aligned}
\min_{\theta_T} \mathcal{L}_{kt}(T) = &\sum_{j \in \mathcal{J}} ||\varphi((A^S_{UDA})^j) - \varphi((A^S_{GA})^j)||_2 \\
&+ \gamma \cdot \sum_{j \in \mathcal{J}} ||\varphi((A^T_{UDA})^j) - \varphi((A^T_{GA})^j)||_2, \\
with, \quad \varphi((A)^j) &= \frac{\sum_{i=1}^{C^j} |A^j_i|^2}{||\sum_{i=1}^{C^j} |A^j_i|^2||_2},
\end{aligned}
\tag{6.9}
$$

where $\mathcal{J}$ indicates the indices of the selected activation layer pairs of GA and MMADT for knowledge transfer, $(A^{S(T)}_{UDA})^j$ and $(A^{S(T)}_{GA})^j$ are 3D activated tensor pairs derived from j-th stage of corresponding encoders, C^j denotes the number of feature channels of j-th activated tensor, $\varphi(\cdot)$ is developed to compute spatial attention map from activated tensor along the feature channel.

Combining the segmentation loss $\mathcal{L}_{seg}$, the depth adversarial training loss $\mathcal{L}_{adv}$, and the knowledge transfer loss $\mathcal{L}_{kt}$, the whole optimization objective of MMADT can be defined as:

$$
\min_{\theta_U} \mathcal{L} = \mathcal{L}_{seg} + \lambda_1 \cdot \mathcal{L}_{adv} + \lambda_2 \cdot \mathcal{L}_{kt}, \tag{6.10}
$$

where λ_1 and λ_2 are loss weighting factors. Integrating all components together, our multi-modal UDA scheme is outlined in Algorithm 2.

Algorithm 2: Multi-modal UDA Scheme

input : Max iterations N, batch size B, hyper-parameters λ_1 and λ_2;
$[\mathcal{X}^S, \mathcal{Y}^S]$: source dataset;
$[\mathcal{X}^T]$: target dataset;
T: initialized semantic segmenter parameterized by θ_U;
D: initialized depth domain discriminator parameterized by θ_D;
GA: pretrained geo-assistant parameterized by θ_{GA};
output: θ_U

training:
for $iter \leftarrow 1$ **to** N **do**
    `// multi-modal sampling`
    Randomly sample $\{x^S_{rgb}, x^S_d, y^S\}^B_{i=1}$ from $[\mathcal{X}^S, \mathcal{Y}^S]$;
    Randomly sample $\{x^T_{rgb}, x^T_d\}^B_{i=1}$ from $[\mathcal{X}^T]$;
    for *each mini-batch* **do**
        Update θ_D of D supervised by $\mathcal{L}_{cls}$ with Adam:
            $\theta_D \leftarrow$ `Adam(BackProp(`$\mathcal{L}_{cls}|\ \theta_U$`)`, θ_D`)`
        Collect gradients of T supervised by $\mathcal{L}_{seg}$, $\mathcal{L}_{adv}$, $\mathcal{L}_{kt}$:
            $\nabla^{seg}_T \leftarrow$ `BackProp(`$\mathcal{L}_{seg}|\theta_U$`)`
            $\nabla^{adv}_T \leftarrow$ `BackProp(`$\mathcal{L}_{adv}|\theta_D$`)`
            $\nabla^{kt}_T \leftarrow$ `BackProp(`$\mathcal{L}_{kt}|\theta_{GA}$`)`
        Update θ_U of T with Adam:
            $\theta_U \leftarrow$ `Adam(`$\nabla^{seg}_T + \lambda_1 \cdot \nabla^{adv}_T + \lambda_2 \cdot \nabla^{kt}_T$, θ_U`)`

testing:
for $idx \leftarrow 1$ **to** `Len(`*test_set*`)` **do**
    `// multi-modal forward pass`
    Get testing sample $\{x_{rgb}, x_d\}_{i=idx}$ and forward pass it to obtain segmentation map;
        $M^{seg} \leftarrow T((x_{rgb}, x_d)|\ \theta_U)$

6.5 Experiments

In this section, we first evaluate the performance of our proposed methods on two benchmark tasks: SYNTHIA-to-Cityscapes and GTA5-to-Cityscapes, and also compare our method with other state-of-the-art methods on these two tasks. Then, we conduct adaptation with a recently published synthetic dataset SELMA [189]. Finally, we illustrate a series of studies to ablate different components and analyze the effectiveness of our multi-modal UDA scheme on different approaches.

6.5.1 Datasets

Cityscapes. The Cityscapes dataset [31] is a large-scale real-world dataset for urban street scene parsing. It contains 5000 finely annotated images captured from 50 cities with 19 semantic object categories, in which 2975 images are used for training, 500 images and 1525 images are used for validation and testing separately. All images are provided with a resolution of 2048×1024. For our multi-modal UDA settings, we use the raw public-available disparity as additional modality and regard the training set without labels as target domain.

SYNTHIA. The SYNTHIA dataset [190] consists of 9400 synthetic image and depth pairs with resolution 1280×760. It adopts the Cityscapes style annotations which share 16 common pixel categories. We regard it as the source domain in SYHTHIA-to-Cityscapes settings.

GTA5. The GTA5 dataset [16] contains 24966 synthetic images with resolution 1280×1052. To accommodate our multi-modal UDA settings, we follow [18] to render the depth map of GTA5 dataset with a pretrained Monodepth2 model [191]. We consider it as the source domain in GTA5-to-Cityscapes settings.

SELMA. The SELMA dataset [189] is a large-scale synthetic dataset, including multiple cameras, and each camera captures 30909 images. In our experiment, we collect the data from desk camera with a resolution of 1280×640 under clear noon weather as the source domain in SELMA-to-Cityscapes settings.

6.5.2 Implementation Details

In all the experiments, we follow a common practice in UDA and resize the resolution of images to 1024×512 for Cityscapes and to 1280×720 for GTA5, and all depth values are normalized into [0,1]. For the segmentation network, we use DeeplabV2 [10] and apply Atrous Spatial Pyramid Pooling (ASPP) with dilated rates of {6, 12}. Note that we utilize ResNet-50 [9] as the feature extractor. For depth adversarial training (ADT), we apply fully-convolutional layers to retain the spatial information, whereas only 4 convolutional layers with kernel 4×4 and a stride of 2 are deployed, in which the channel numbers are {128,256,512,1}, respectively. In addition, we use a max-pooling operator with kernel 2×2 and a stride of 2 to down-sample the feature map before the convolutional layers. For Geo-Assistant (GA), we adopt the same encoder structure as in the segmentation network for feature embedding and a straightforward

depth decoder structure for depth reconstruction. Specifically, the depth decoder consists of 3 convolutional groups plus a regression head. Each convolutional group contains 2 convolutional layers with kernel 3×3 and a stride of 1. The channel numbers of each group are {2048, 256, 128}, respectively. After each convolution group, the feature map is interpolated by a sampling rate of 2 until the output size is consistent with the original image.

All experiments are run on a single RTX3090 GPU with 24GB memory. We optimize the whole networks by Adam [188] with a weight decay of 5×10^{-4}. We empirically set an initial learning rate of 6×10^{-5} for encoder and 6×10^{-4} for all decoders. We linearly warm up the learning rate with 1500 iterations during training, then linear decay afterward. For self-supervised GA, we experimentally set α_1 and α_2 to 1.0. For UDA training, the hyper-parameters λ_1 is set to 1.0 for SYNTHIA and GTA5 to Cityscapes adaptation and set to 1.5 for SELMA to Cityscapes adaptation. Then, γ is set to 1.0, and λ_2 is set to 0.001. In addition, all experiments adopt Rare Class Sampling (RCS) and Thing-Class ImageNet Feature Distance (FD) strategies, as described in [192], to stabilize and speed up the convergence. During training, the inputs are randomly cropped to 512×512 with a mini-batch size of 2 for both source and target domains. We train self-supervised multi-modal GA and MMADT for 40k iterations for SYNTHIA-to-Cityscapes setup, and for 80k iterations for GTA5-to-Cityscapes and SELMA-to-Cityscapes setup.

Baselines. To comprehensively evaluate our method, we implement two baseline models, namely Baseline and Baseline-MM. More concretely, the Baseline only takes RGB images as input, while the Baseline-MM holds concatenated RGB and depth pairs as input. Thus, the Baseline-MM can be seen as a multi-modal version of the Baseline. The two baseline models use the same segmentation network as our MMADT, i.e., DeeplabV2 with ResNet-50, and follow the same training strategy.

6.6 Results

We first evaluate the semantic segmentation performance in terms of mean intersection over union (mIoU) over two commonly studied UDA tasks. In table 6.1, we compare our approach with the state-of-the-art methods over 16 classes on SYNTHIA-to-Cityscapes setup, '$\surd$' and '$\times$' reflect whether multi-modal data, i.e., RGB-D, is used during training or testing. Compared with those methods that only utilize RGB information, our MMADT achieves better results. Referring to the results of Baseline-MM, using concatenated RGB-D data directly in both training and testing drives performance degradation. It reflects that introducing depth input will raise an additional domain gap which offsets the benefits of geometric clues and increases the difficulty of the UDA task. Then, thanks to the proposed DFB, DAT, and self-supervised multi-modal GA, MMADT can mitigate the adverse effects of raw depth input and explore the additional geometric cues. As shown in table 6.1, MMADT far surpasses the Baseline and Baseline-MM by 3.7% and 5.0% over 16 classes, 4.0% and 5.5% over 13 classes.

We also report the mIoU over 6 moving objects classes, namely 'person', ' rider', 'car', 'bus', 'moto.', 'bike', marked with $\diamond$ in table 6.1, which are more salient in the depth image. The Baseline-MM yields 50.6% mIoU, while our MMADT

Table 6.1: Comparisons of our MMADT with the state-of-the-art depth-aware based UDA methods over SYNTHIA-to-Cityscapes set up. Our method far exceeds other UDA methods by a large margin. mIoU* denotes performance over 13 classes excluding those marked with $\star$ and mIoU$^\diamond$ denotes performance over 6 moving classes marked with $\diamond$.

Method	RGB-D		Backbone	road	s.walk	build.	wall*	fence*	pole*	light	sign	veget.
	Train	Test										
SPIGAN [166]	√	×	Vgg19	71.1	29.8	71.4	3.7	0.3	33.2	6.4	15.6	81.2
GIO-Ada [100]	√	×	Res101	78.3	29.2	76.9	11.4	0.3	26.5	10.8	17.2	81.7
DADA [19]	√	×	Res101	89.2	44.8	81.4	6.8	0.3	26.2	8.6	11.1	81.8
CTRL [20]	√	×	Res101	86.4	42.5	80.4	20.0	1.0	27.7	10.5	13.3	80.6
CorDA [18]	√	×	Res101	**93.3**	**61.6**	**85.3**	19.6	**5.1**	37.8	36.6	42.8	84.9
Baseline[192]	×	×	Res50	76.4	34.5	84.9	21.3	2.3	37.6	44.0	**50.1**	**84.9**
Baseline-MM	√	√	Res50	84.0	44.0	84.2	19.3	3.3	35.6	**44.6**	47.2	84.3
MMADT (Ours)	√	√	Res50	90.0	57.3	84.5	**23.8**	4.6	**39.4**	42.5	48.5	84.2

sky	person$^\diamond$	rider$^\diamond$	car$^\diamond$	bus$^\diamond$	moto.$^\diamond$	bike$^\diamond$	mIoU$^\diamond$	mIoU*	mIoU
78.9	52.7	13.1	75.9	25.5	10.0	20.5	33.0	42.4	36.8
81.9	45.8	15.4	68.0	15.9	7.5	30.4	30.5	43.0	37.3
84	54.7	19.3	79.7	40.7	14.0	38.8	41.2	49.8	42.6
82.6	61.0	23.7	81.8	42.9	21.0	44.7	45.9	51.5	45.0
90.4	69.7	41.8	**85.6**	38.4	32.6	53.9	53.7	62.8	55.0
86.1	**70.1**	40.3	83.3	45.3	40.7	53.1	55.5	61.1	53.4
83.3	66.1	40.9	81.3	41.6	25.1	48.7	50.6	59.6	52.1
85.1	68.8	**43.9**	82.0	**49.9**	**47.2**	**62.3**	**59.0**	**65.1**	**57.1**

Table 6.2: Comparisons of our MMADT with the state-of-the-art UDA methods over GTA5-to-Cityscapes set up. Our method outperforms other UDA methods by a large margin. mIoU$^\diamond$ denotes performance over 6 moving classes marked with $\diamond$.

Method	RGB-D		Backbone	road	s.walk	build.	wall	fence	pole	light	sign	veget.
	Train	Test										
ADVENT [168]	×	×	Res101	87.6	21.4	82.0	34.8	26.2	28.5	35.6	23.0	84.5
CBST [171]	×	×	Res101	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9
DACS [174]	×	×	Res101	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0
ProDA [193]	×	×	Res101	87.8	56.0	79.7	**46.3**	**44.8**	**45.6**	**53.5**	53.5	**88.6**
CorDA [18]	√	×	Res101	94.7	63.1	87.6	30.7	40.6	40.2	47.8	51.6	87.6
Baseline [192]	×	×	Res50	94.5	66.1	86.8	31.9	30.3	35.6	45.3	53.3	87.1
Baseline-MM	√	√	Res50	92.8	64.2	86.5	32.6	26.0	37.3	48.3	53.7	87.4
MMADT (Ours)	√	√	Res50	**96.0**	**73.1**	**88.3**	41.0	**33.3**	41.3	51.4	**62.0**	88.2

Terrain	sky	person$^\diamond$	rider$^\diamond$	car$^\diamond$	trunk$^\diamond$	bus$^\diamond$	train$^\diamond$	moto.$^\diamond$	bike$^\diamond$	mIoU$^\diamond$	mIoU
35.1	76.2	58.6	30.7	84.8	34.2	43.4	0.4	28.4	35.3	39.5	44.8
34.2	80.9	53.1	24.0	82.7	30.3	35.9	**16.0**	25.9	42.8	38.8	45.9
44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.	27.3	34.0	43.1	52.1
45.2	82.1	**70.7**	39.2	88.8	45.5	**59.4**	1.0	48.9	56.4	51.2	57.5
47.0	89.7	66.7	35.9	**90.2**	**48.9**	57.5	0.	39.8	56.0	49.4	56.6
43.8	89.2	67.6	38.7	86.0	45.2	46.3	0.2	43.9	56.4	48.0	55.2
44.3	90.8	63.9	37.6	81.7	42.0	48.3	0.	37.5	57.2	46.0	54.3
45.9	**91.6**	67.9	**41.4**	88.8	48.4	58.3	0.4	**52.5**	**61.7**	**52.4**	**59.6**

Table 6.3: Comparisons of our MMADT with the baselines over SELMA-to-Cityscapes set up. Our method outperforms our baseline methods by a large margin. mIoU$^{\diamond}$ denotes performance over 6 moving classes marked with $\diamond$.

Method	RGB-D		Backbone	road	s.walk	build.	wall	fence	pole	light	sign	veget.
	Train	Test										
Source-only [189]	×	×	Res101	88.5	52.1	81.9	21.8	14.3	29.2	16.4	31.2	83.0
Baseline [192]	×	×	Res50	**92.2**	55.8	87.2	21.7	25.5	**41.7**	**48.2**	56.5	**85.9**
Baseline-MM	√	√	Res50	92.6	60.9	84.9	11.9	20.6	33.2	43.8	53.1	85.6
MMADT (Ours)	√	√	Res50	91.4	**60.5**	**87.5**	**37.2**	**29.3**	37.6	47.1	**58.7**	82.3

Terrain	sky	person$^{\diamond}$	rider$^{\diamond}$	car$^{\diamond}$	trunk$^{\diamond}$	bus$^{\diamond}$	train$^{\diamond}$	moto.$^{\diamond}$	bike$^{\diamond}$	mIoU$^{\diamond}$	mIoU
26.0	78.7	54.7	27.6	69.2	15.5	9.1	21.9	32.0	35.3	33.2	41.5
35.7	89.5	67.4	42.3	80.9	24.0	**63.3**	**23.3**	50.5	56.0	50.2	55.1
33.1	87.4	63.1	36.3	80.5	23.5	41.0	18.2	32.8	56.7	44.0	50.5
33.1	**91.0**	**68.7**	**42.8**	**86.2**	**35.0**	62.5	12.8	**58.9**	**60.5**	**53.4**	**57.0**

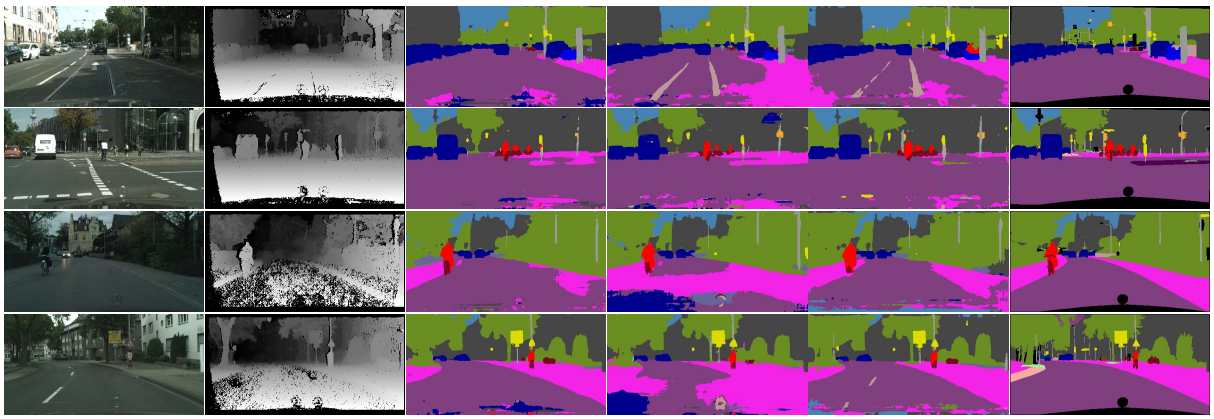


Figure 6.3: Semantic segmentation qualitative results on SYNTHIA-to-Cityscapes set up. Column (a) and (b) show the RGB and depth inputs, segmentation map of MMADT, Baseline and Baseline-MM are illustrated in column (c) - (e), column (f) presents the ground truth segmentation map with void class (black pixels).

yields 59.0% mIoU, which achieves a large margin of 8.4% absolute improvement. This further demonstrates the effectiveness of our approach when dealing with multi-modal data, i.e., RGB-D. In addition, our method yields a 2.1% mIoU improvement over the current best depth-aware method [18] on 16 classes and a 5.3% mIoU raise on the 6 moving objects categories.

For the GTA5-to-Cityscapes setup, we observe a consistent improvement of mIoU over 19 classes, as reported in Table 6.2. Note that we use a coarse depth image inferred from RGB as input, which will affect the performanfourthce of our MMADT to some extent. But our method still outperforms other state-of-the-art methods and its baselines. Specifically, our approach obtains 59.6% mIoU, better than the 54.3% and 55.2% mIoU of our baselines and the 57.5% mIoU of the state-of-the-art method [193]. Then, to further illustrate the effectiveness of our method, we perform an adaptation from the lately published SELMA synthetic dataset. As shown in Table 6.3, our method significantly outperforms the RGB-based baseline, as well as the simple RGB-D-based baseline.

In addition, we provide some visual examples of segmentation results in Figure 6.3. We can see that facilitated by the geometric prior contained in the depth image, the prediction quality is highly improved on easily twisted classes and small-scale objects, such as road vs. sidewalk and bike vs. rider.

To better explain the intuition of the proposed multi-modal UDA scheme, we compare the features learned by our baselines and MMADT. Specifically, we map the feature space of the last hidden layer to 2D space by t-SNE [194] under SYNTHIA-to-Cityscapes adaptation setup. As shown in Figure 6.4, the upper and lower parts present the features space of 16 classes and 6 moving object classes in the Cityscapes validation dataset, respectively. We can see that for the object classes visible in the depth map, the multi-modal information helps feature clustering. Yet, directly using the raw depth produces a large clustering radius, which leads to confusion about decision boundaries. It further explains the reason for the performance degradation of Baseline-MM. Compared with the baselines, our MMADT yields a smaller clustering radius as well as more distinct category boundaries, which reflects that our

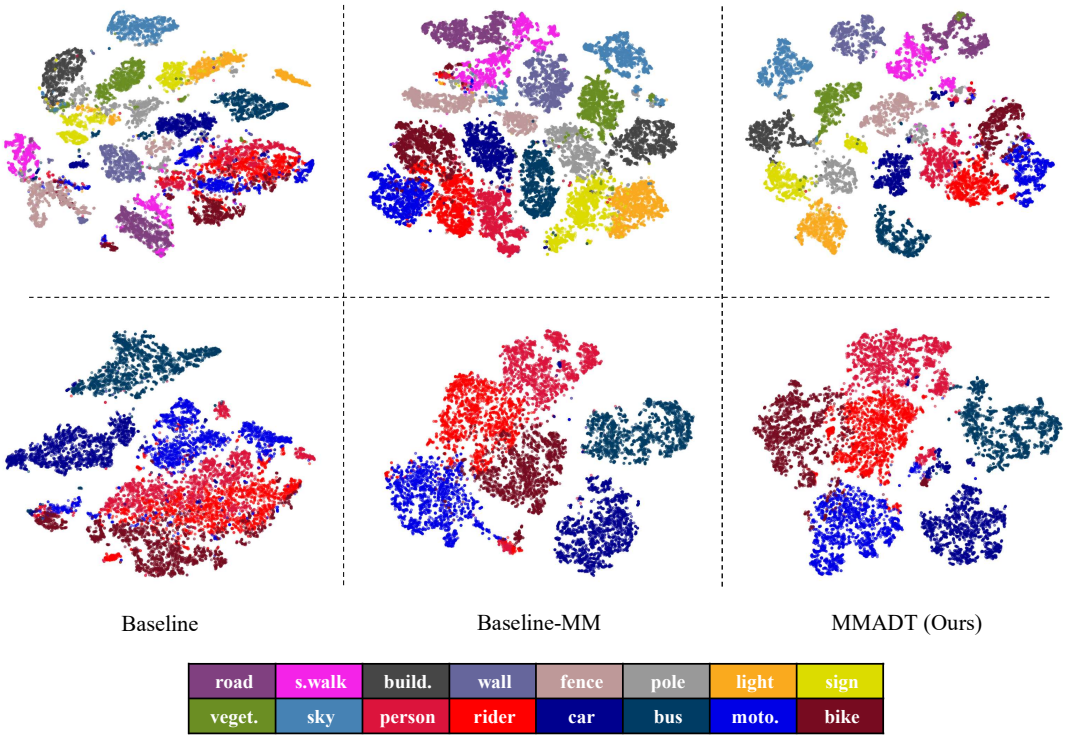


Figure 6.4: t-SNE visualization of feature space on Cityscapes validation set. Upper: visualization of 16 classes. Lower: visualization of 6 moving object classes, i.e. 'person', 'rider', 'car', 'bus', 'motocycle', 'bike'. Each color represents one specific semantic class.

Table 6.4: The effects of loss weight factors in self-supervised multi-modal GA on MMADT. Here we only report the mIoU over 16 classes on SYNTHIA-to-Cityscapes setup.

α_2	0.01	0.1	0.5	1.0	1.5	2.0	3.0
mIoU	55.6	56.1	56.5	**57.1**	56.2	55.8	55.6

proposed method can effectively take advantage of different modalities to enhance the performance of the UDA model.

6.6.1 Geo-Assistant Analysis

Self-supervised multi-modal GA is a key component of our MMADT, thus we study the impact of the weight factors α_1 and α_2 in Algorithm 1. Basically, α_1 and α_2 are specified to balance the contribution of the source and target domain. Particularly, we set α_1 as 1.0 and produce experiments with different α_2 values. When $\alpha_2 < 1.0$, GA is dominated by the source data, while when $\alpha_2 > 1.0$ the target data dominates. As shown in Table 6.4, the best performance is achieved when α_2 equals 1.0. This result is very intuitive because in terms of adapting data from the source domain to the target domain, we expect the GA to generalize well on the target domain, so it makes sense to let the target domain dominate the GA, while an overweight of the target domain would cause the noise of raw depth to mislead the learning process. Hence, we empirically set $\alpha_1 = 1.0$ and $\alpha_2 = 1.0$ in our experiments.

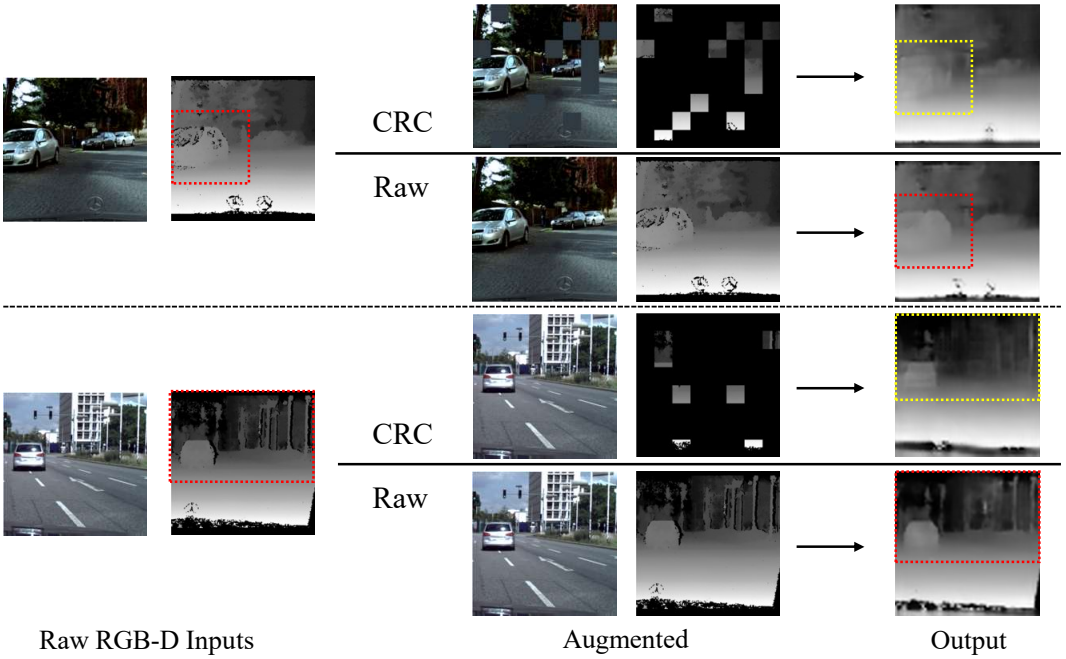


Figure 6.5: Visualization of depth estimation results of self-supervised Geo-Assistant on Cityscapes validation set. CRC and Raw mean the input data processed w/ and w/o CRC data augmentation.

Table 6.5: Ablation study of the components of our MMADT on SYNTHIA-to-Cityscapes set up. ● and ○ indicate activated and inactivated components, respectively. ◉̂ means pre-training GA without CRC strategy.

Method	DFB	DAT	GA	mIoU$^{\diamond}$	mIoU*	mIoU
Baseline-MM	○	○	○	50.1	58.6	51.4
MMADT	●	○	○	55.3	60.5	53.2
MMADT	●	●	○	57.0	64.1	55.8
MMADT	●	●	●	**59.0**	**65.1**	**57.1**
MMADT	●	●	◉̂	57.2	63.9	56.2

In addition, as we can see in Figure 6.5, the trained GA is capable of reconstructing depth from CRC augmented or raw data, which demonstrates that our GA can be aware of the depth clues and fuse them with the RGB features. Moreover, we observe that the trained GA is robust to noise, and the depth maps reconstructed from the target domain are closer in style to the source domain, which is helpful in reducing the domain gap for UDA tasks.

6.6.2 Ablation Studies

6.6.2.1 The effectiveness of DFB and DAT

For a fair comparison, we implement a multi-modal baseline named Baseline-MM. As described in Section 6.6, Baseline-MM blends RGB and depth images by direct concatenation. Instead of fusing all depth information indiscriminately, our DFB combines the calibrated depth features, thus eliminating the side effect of directly using raw depth to some extent. Moreover, applying DAT to bridge the appearance gap of depth images between the source

Table 6.6: The effect of distilled layers of GA to MMADT. l_1 to l_4 refer to the different encoding layers in the encoder. • and ○ indicate activated and inactivated distilled layers, respectively.

Method	l_1	l_2	l_3	l_4	mIoU
MMADT	•	•	•	•	55.9
MMADT	•	•	•	○	**57.1**
MMADT	•	•	○	•	55.8
MMADT	•	○	•	•	56.3
MMADT	○	•	•	•	55.5

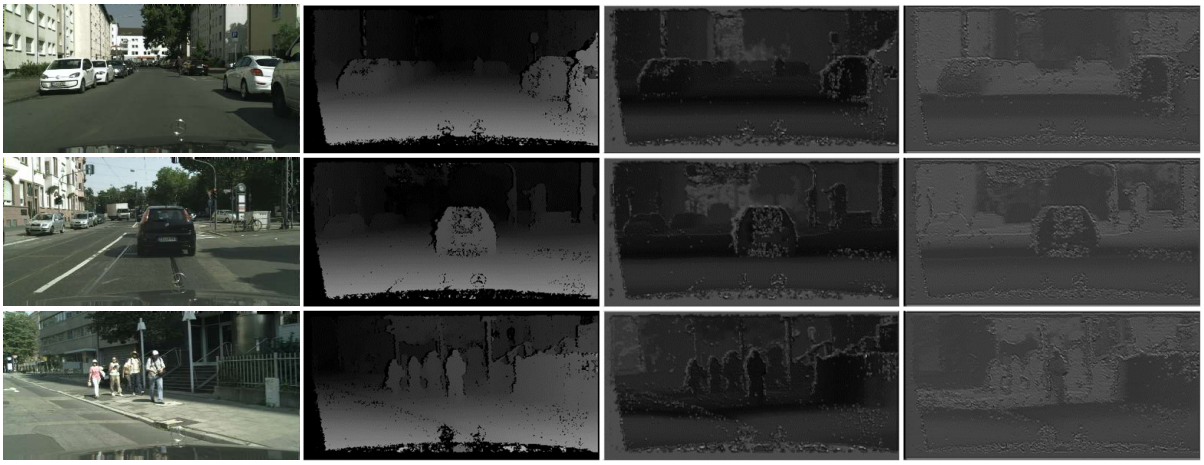
and target domain can further enhance the performance of our model, as shown in the top three rows of Table 6.5. As shown in Table 6.5, the different components of our method gradually improve the segmentation performance.

In figure 6.6, we visualize the depth features before and after calibration as detailed in Section 6.4.1. We can see that the calibrated depth features focus more on meaningful things such as people, cars, plants, etc., which can be used as a complement to the RGB features. Therefore, DFB and DAT are critical to executing effective multi-modal fusion.

6.6.2.2 The effect of CRC and distilled layers in Geo-Assistant

The fourth row of Table 6.5 shows that knowledge transfer from GA to MMADT can further improve the performance. Then, to show the effectiveness of CRC strategy, we explicitly bypass CRC module during pre-training GA and keep the rest of our designs unchanged. We obtained 56.2% of mIoU, a decrease of 0.9 % compared to our result with CRC. Furthermore, we observe a drop of 1.8% mIoU for those moving categories salient in the depth images. It further highlights that using CRC allows the model to identify valid features in the depth image to complement RGB information, while these features are typically prone to be overlooked in model training without CRC. The results are shown in the last row of Table 6.5.

We additionally compare the impact of the choice of distilled layers. As listed in Table 6.6, we perform extensive experiments with different layer choices in the SYNTHIA-to-Cityscapes setup. Since it is expected to narrow the attention distribution by transferring task-independent functionalities from the GA network to the MMADT, we select at least three feature extraction layers for knowledge transfer in order to ensure that our MMADT can learn universal and valuable attributes from pretrained GA. In table 6.6, l_1 to l_4 denote the four Res-layers in ResNet [9]. We can see that distilling the knowledge of l_4 leads to severe performance degradation, suggesting that l_4 may contain more task-specific knowledge, which will hinder the optimization of our model. Therefore, we empirically determine to transfer the features of l_1 to l_3 from GA to MMADT.



(a) Images (b) Depth (c) DF (d) DF Calibrated

Figure 6.6: Visualization of depth features (DF) before vs. after calibration of DFB. Column (a) and (b) present the original RGB and depth input. Column(c) illustrates the depth features w/o calibration, and column (d) presents the calibrated depth features that will be combined with color features.

Table 6.7: Flexibility analysis on SYNTHIA-to-Cityscapes setup. Our multi-modal UDA method can be smoothly ported to existing UDA models with consistent performance progress.

Method	RGD-D		Backbone	mIoU*	mIoU
	Train	Test			
DACS [174]	×	×	Res50	54.1	47.3
DACS-MM (Ours)	√	√	Res50	**59.6**	**52.2**
ADVENT [168]	×	×	Res50	48.0	41.7
ADVENT-MM (Ours)	√	√	Res50	**50.6**	**44.1**

6.6.3 Flexibility Analysis

Our multi-modal UDA method is equipped with three designs, namely DFB, DAT, and GA, which can be easily ported to existing UDA methods as a plug-in. We adapt two representative works, namely DACS [174] and ADVENT [168], to evaluate the flexibility, in which DACS is a self-training-related approach and ADVENT is an adversarial-training-related approach. To this end, a multi-modal version of DACS and ADVENT are implemented, namely DACS-MM and ADVENT-MM, respectively. Specifically, we plug the DFB in the encoder for connecting depth information and insert DAT ($\mathcal{L}_{adv}$) and GA ($\mathcal{L}_{kt}$) to optimize the multi-modal UDA model. In addition, we leave the hyper-parameter settings unchanged and train the model exactly as the original implementation. Note that we use ResNet-50 as the backbone of all models for a fair comparison. As shown in Table 6.7, our multi-modal method can be seamlessly applied to existing UDA methods and achieves significant performance gains on both self-training-related and adversarial-training-related methods.

6.7 Summary

In this chapter, we present a novel multi-modal UDA framework for semantic segmentation, which aims at leveraging additional information to boost adaptation performance. To do this, we treat the depth image as an auxiliary input and train the model under a multi-modal learning paradigm, in which we encounter two challenges. Firstly, discrepancies among the domains of auxiliary modality further exacerbate the domain gap. Secondly, the features between different modalities are not necessarily perfectly aligned, especially in the target domain. To address these challenges, we propose a multi-modal network named MMADT, composed of three key designs, i.e., Depth fusion block (DFB), Depth adversarial training (DAT), and Geo-assistant (GA). The proper application of these components in a multi-modal network helps the model mitigate the discrepancies between the same modalities in different domains and the alignment between different modalities in same domains. Additionally, the accretion of modality-specific information facilitates the UDA model to learn consistent feature representation, thereby improving the generalization ability of the model over different domains. To the best of our knowledge, this is the first proposed work to solve the UDA problem under the multi-modal learning paradigm. Besides, our multi-modal UDA training strategy can also be freely ported to existing UDA models. Extensive experiments illustrate that additional modality can effectively enhance the model's parsing ability and resist domain shifts. Our results far exceed the baselines and start-of-the-art UDA methods.

In this work, we formulate a UDA learning strategy under the multi-modal learning paradigm. We find that the performance of UDA can be remarkably enhanced by utilizing multi-modal information by exploring the complementary nature of different modality information, as well as developing more efficient fusion methods to improve performance.

Chapter 7

Conclusion and perspective

7.1 Conclusion

In this thesis, we are interested in exploring multi-modal visual data fusion methods to enhance the accuracy and efficiency of outdoor scene analysis. Our primary concern is how to effectively utilize multi-modal visual data, such as color images, infrared images, and depth images, as well as how to fuse these visual data to provide a more comprehensive understanding of the environment. We select two representative computer vision tasks, namely semantic segmentation and object detection, as our points of entry for investigating and validating various multi-modal visual data fusion methods. To this end, in Chapter 2, we initially introduce the fundamental concepts and existing mainstream methods for semantic segmentation and object detection and then elaborate on the multi-modal vision data fusion framework and techniques. Furthermore, we investigate unsupervised domain adaptation as a strategy to address changes in data distribution, specifically in scenarios where training data is scarce in real-world settings. At the end of Chapter 2, we provide a comprehensive review and analysis of methods related to unsupervised domain adaptation with a particular focus on the application of multiple modalities in it. In Chapters 3 and 4, we explore both single-modal and multi-modal semantic segmentation models within the encoder-decoder framework. In Chapter 5, we delve further into how to tackle object detection issues under challenging weather conditions, specifically low lighting, using multi-modal fusion techniques. Lastly, we shift our focus to unsupervised domain adaptation research in Chapter 6 and discuss how to leverage multi-modal data fusion to minimize distribution discrepancies between domains. Our primary efforts can be summarized as follows:

In our research, we initially study semantic segmentation methods based on a single modality, i.e., RGB images, and seek to enhance the performance of existing methods by optimizing training strategies and model architectures. To achieve this, we proposed a general two-branch decoder paradigm along with a boundary-enhanced loss strategy. The two decoders can adaptively learn complementary information without explicitly designating specific learning elements. Experiments suggest that introducing additional boundary information in the loss function and making

two branches compete during training can improve the training efficiency of the model to some extent.

Subsequently, we propose an RGB-D fusion scheme based on additive attention, where the Depth map is regarded as an auxiliary modality supplying additional geometric cues. Our goal is to utilize additive attention to replace the complex matrix computation process in the original self-attention and thus solve the prohibitive cost issues linked to self-attention-related multi-modal fusion methods. To this end, we introduce a lightweight fusion network called HCFNet. This network is capable of retaining local details while probing the long-range dependencies of multi-modal information, thereby extracting complementary information from different modalities. The efficacy of our proposed method is confirmed through testing on both indoor and outdoor datasets.

In addition, considering the complexity of scene perception under low light conditions, we capitalize on the complementary information between thermal infrared and visible light images to enhance the perceptual capability of the system to its surroundings. Given the flexibility of RGB-T fusion, we introduce a lightweight cross-fusion module named Channel-Patch Cross Fusion (CPCF). This module leverages cross-attention at both the channel and patch levels to encourage mutual correction between different modalities while maintaining their unique properties. Extensive experiments demonstrate that the proposed method outperforms others on several publicly available multi-modal datasets. Besides, it can be extended to different types of detectors, further showcasing its robustness and generalizability.

Finally, we explore how to leverage multi-modal information to assist the model in learning domain-independent feature representations in the unsupervised domain adaptation setup, thereby reducing the gap between different domains while expanding the inter-class distance in the semantic feature space. In line with this, we propose a new multi-modal-based unsupervised domain adaptation method called MMADT, which aims to fully utilize the input RGB and depth information in semantic segmentation tasks. We design a Depth Fusion Block (DFB) and a Depth Adversarial Training (DAT) strategy to narrow the depth discrepancy between the source and target domains. Then, we propose a self-supervised multi-modal depth estimation assistant network called Geo-Assistant (GA) to align the RGB and depth at the feature spaces. We observe significant performance improvements in multiple synthetic-to-real adaptation benchmarks.

7.2 Perspective

In this thesis, we aim to harness multi-modal information to bolster the perception capabilities of models. To achieve this goal, we have introduced a suite of multi-modal fusion schemes designed to address various computer vision tasks, with a focus on fusion efficiency and model complexity. Despite some positive progress, many technical challenges and unexplored areas in multi-modal learning still need to be explored. In light of the contributions of this thesis, we will discuss potential future research directions and provide a projection on the evolution of multi-modal learning based on our current knowledge and understanding.

We have proposed two lightweight fusion schemes based on different attention mechanisms for RGB-D and RGB-T data fusion and exhibited impressive performance across multiple public datasets. However, our model was trained and tested on strictly aligned image pairs, which limits its applicability in real-world scenarios to a certain extent. For instance, in autonomous driving or drone systems, RGB and depth or infrared images are captured by different cameras, which may suffer from issues of camera displacement and rotation, making image alignment challenging. Therefore, multi-modal fusion on non-aligned image pairs represents a promising avenue for future research. Furthermore, another challenge brought about by multimodality is the issue of missing modalities and signal noise. Future research needs to consider how to ensure the stability and robustness of the model under these circumstances. For diverse weather conditions, we have explored using multi-modal cues to enhance the model's performance under low-light conditions. Nevertheless, we have yet to delve into other challenging environmental conditions like fog, rain, and snow. In addition, viewing from a broader perspective, with the advent of a wider variety of sensor technologies and data collection methods, richer multi-modal data, such as sonar, radar, and event, will provide signal attributes different from traditional visual sensors. Consequently, how to integrate these signals to meet the challenges posed by various weather conditions is also a direction to be explored. On the other hand, in unsupervised domain adaptation, we have attempted to minimize the distribution difference between domains by employing RGB-D data. While we have achieved significant performance improvements in some benchmark tests, the efficacy of this method in real-world scenarios remains insufficiently verified. Especially when there is a significant discrepancy in data distributions, or the model needs to transfer between multiple source and target domains, the effective use of multi-modal data to bridge inter-domain gaps still requires further exploration. Besides, the effects of different modalities on model adaptation may vary in different scenarios.

The potential of multi-modal data fusion is huge, and the overarching goal in this field is to design a universal framework that can accommodate various modalities while gracefully handling issues related to modality absence and noise. Although recent research [195] has proposed a Transformer-based fusion architecture, which fused RGB images with Depth, Thermal, or Event images, has demonstrated promising results across different modalities, it still relies on aligned data, and the mechanics of fusion remain in the exploratory phase. Hence, there is still significant research space in exploring general representations of different modalities and feature fusion approaches based on different attention mechanisms. In addition, with the innovation of deep model architectures and the improvements in hardware computational power, another promising research direction is the knowledge transfer from large-scale language pre-training model [196] or language-image pre-training model [197] to downstream multi-modal data fusion frameworks. Recently, several studies [198, 199] have demonstrated the advantages of language pre-training models in understanding and generating latent semantic relations in language, which provides new perspectives for improving the comprehension of multi-modal systems. For instance, using language pre-training models for correlating and reasoning the visual and textual information in multi-modal data could potentially contribute to improving multi-modal fusion results and scene analysis capabilities.

To conclude, the multi-modal visual data fusion methods involved in this study merely represent the tip of the iceberg. Future research is anticipated to encompass broader and deeper content, with the aim of achieving greater breakthroughs in the fields of computer vision and machine learning.

Bibliography

[1] Sijie Hu., Fabien Bonardi., Samia Bouchafa., and Désiré Sidibé. A general two-branch decoder architecture for improving encoder-decoder image segmentation models. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP, (VISIGRAPP 2022)*, pages 374–381. INSTICC, SciTePress, 2022.

[2] Sijie Hu, Fabien Bonardi, Samia Bouchafa, and Désiré Sidibé. A hybrid multi-modal visual data cross fusion network for indoor and outdoor scene segmentation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2539–2545, 2022.

[3] Sijie Hu, Fabien Bonardi, Samia Bouchafa, and Désiré Sidibé. Multi-modal unsupervised domain adaptation for semantic image segmentation. *Pattern Recognition*, 137:109299, 2023.

[4] Michael E. Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13:290, 1982.

[5] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[8] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[11] Ross B. Girshick. Fast r-cnn. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[12] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and K. Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2019.

[13] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10745–10755, 2021.

[14] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image Vis. Comput.*, 105:104042, 2021.

[15] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[16] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[17] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016.

[18] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021.

[19] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7364–7373, 2019.

[20] Suman Saha, Anton Obukhov, Danda Pani Paudel, Menelaos Kanakis, Yuhua Chen, Stamatios Georgoulis, and Luc Van Gool. Learning to relate depth and semantics for unsupervised domain adaptation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8193–8203, 2021.

[21] Dana Lahat, T. Adalı, and Christian Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103:1449–1477, 2015.

[22] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:1964–1980, 2019.

[23] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep multimodal learning for audio-visual speech recognition. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134, 2015.

[24] Saurabh Gupta, Ross B. Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. *ArXiv*, abs/1407.5736, 2014.

[25] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4:2576–2583, 2019.

[26] Yongtao Zhang, Zhishuai Yin, Linzhen Nie, and Song Huang. Attention based multi-layer fusion of multispectral images for pedestrian detection. *IEEE Access*, 8:165071–165084, 2020.

[27] Gongyang Li, Zhi Liu, and Haibin Ling. Icnet: Information conversion network for rgb-d based salient object detection. *IEEE Transactions on Image Processing*, 29:4873–4884, 2020.

[28] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[29] Adam Dlesk, Karel Vach, and Karel Pavelka. Photogrammetric co-processing of thermal infrared images and rgb images. *Sensors*, 22(4), 2022.

[30] Yinglong Dai, Zheng Yan, Jiangchang Cheng, Xiaojun Duan, and Guojun Wang. Analysis of multimodal data fusion from an information theory perspective. *Information Sciences*, 623:164–183, 2023.

[31] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[32] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[33] Garrett Wilson and Diane J. Cook. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.*, 11(5), jul 2020.

[34] Saeid Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and G. Hamarneh. Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 54:137–178, 2019.

[35] Çağrı Kaymak and Ayşegül Uçar. A brief survey and an application of semantic image segmentation for autonomous driving. In *Handbook of Deep Learning Applications*, 2018.

[36] Xian Tao, Dapeng Zhang, Wenzhi Ma, Xilong Liu, and De Xu. Automatic metallic surface defect detection and recognition with convolutional neural networks. *Applied Sciences*, 2018.

[37] Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321–348, 2019.

[38] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2015.

[41] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[42] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[43] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[45] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[46] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[47] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5168–5177, 2016.

[48] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[49] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *European Conference on Computer Vision*, pages 489–506. Springer, 2020.

[50] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2013.

[51] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104:154–171, 2013.

[52] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[53] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2016.

[54] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2015.

[55] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2016.

[56] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *ArXiv*, abs/1804.02767, 2018.

[57] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *ArXiv*, abs/2004.10934, 2020.

[58] W. Liu, Dragomir Anguelov, D. Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, 2015.

[59] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327, 2017.

[60] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *ArXiv*, abs/2207.02696, 2022.

[61] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. *International Journal of Computer Vision*, 128:642–656, 2018.

[62] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9626–9635, 2019.

[63] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *ArXiv*, abs/2107.08430, 2021.

[64] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge J. Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2017.

[65] Lei Liu, Zongxu Pan, and Bin Lei. Learning a rotation invariant detector with rotatable bounding box. *ArXiv*, abs/1711.09405, 2017.

[66] Jiaming Han, Jian Ding, Jie Li, and Guisong Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2020.

[67] Xingjia Pan, Yuqiang Ren, Kekai Sheng, Weiming Dong, Haolei Yuan, Xiao-Wei Guo, Chongyang Ma, and Changsheng Xu. Dynamic refinement network for oriented and densely packed object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11204–11213, 2020.

[68] Yifei Zhang, Desire Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image Vis. Comput.*, 105:104042, 2020.

[69] Kuan Liu, Yanen Li, N. Xu, and P. Natarajan. Learn to combine modalities in multimodal deep learning. *ArXiv*, abs/1805.11730, 2018.

[70] Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7088–7097, 2021.

[71] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013.

[72] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas. Multispectral deep neural networks for pedestrian detection. *ArXiv*, abs/1611.02644, 2016.

[73] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1475–1483, 2017.

[74] Hermann Blum, Abel Gawel, Roland Y. Siegwart, and César Cadena. Modular sensor fusion for semantic segmentation. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3670–3677, 2018.

[75] Abhinav Valada, Gabriel L. Oliveira, Thomas Brox, and Wolfram Burgard. Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In *International Symposium on Experimental Robotics*, 2016.

[76] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4644–4651, 2017.

[77] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *ArXiv*, abs/1803.05347, 2018.

[78] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer, 2016.

[79] Shang-Wei Hung and Shao-Yuan Lo. Incorporating luminance, depth and color information by a fusion-based network for semantic segmentation. *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2374–2378, 2018.

[80] Liuyuan Deng, Ming Yang, Tianyi Li, Yuesheng He, and Chunxiang Wang. Rfbnet: Deep multimodal networks with residual fusion blocks for rgb-d semantic segmentation. *ArXiv*, abs/1907.00135, 2019.

[81] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, pages 1–47, 2019.

[82] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019.

[83] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[84] Jiangtao Xu, Kaige Lu, and Han Wang. Attention fusion network for multi-spectral semantic segmentation. *Pattern Recognit. Lett.*, 146:179–184, 2021.

[85] Fang Qingyun and Wang Zhaokui. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, 2022.

[86] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. Unsupervised domain adaptation in semantic segmentation: a review. *ArXiv*, abs/2005.10876, 2020.

[87] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alex Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[88] Eric Tzeng, Judy Hoffman, N. Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *ArXiv*, abs/1412.3474, 2014.

[89] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. *ArXiv*, abs/1502.02791, 2015.

[90] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshops*, 2016.

[91] Kuniaki Saito, Y. Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, 2017.

[92] Chaoqi Chen, Weiping Xie, Tingyang Xu, Wen bing Huang, Yu Rong, Xinghao Ding, Yue Huang, and Junzhou Huang. Progressive feature alignment for unsupervised domain adaptation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 627–636, 2018.

[93] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4888–4897, 2019.

[94] Woonggi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7346–7354, 2019.

[95] Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. In *J. Mach. Learn. Res.*, 2016.

[96] Mingsheng Long, Hanhua Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*, 2016.

[97] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017.

[98] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018.

[99] Changchun Zhang, Qingjie Zhao, and Yu Wang. Hybrid adversarial network for unsupervised domain adaptation. *Inf. Sci.*, 514:44–55, 2020.

[100] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1841–1850, 2019.

[101] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[102] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[103] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9167–9176, 2019.

[104] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[105] Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Semeda: Enhancing segmentation precision with semantic edge aware loss. *Pattern Recognition*, 108:107557, 2020.

[106] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[107] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13065–13074, 2020.

[108] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[109] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. *arXiv preprint arXiv:2012.15840*, 2020.

[110] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[111] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5229–5238, 2019.

[112] Qin Huang, Chunyang Xia, Chihao Wu, Siyang Li, Ye Wang, Yuhang Song, and C-C Jay Kuo. Semantic segmentation with reverse attention. *arXiv preprint arXiv:1707.06426*, 2017.

[113] Abhinav Valada, Gabriel Oliveira, Thomas Brox, and Wolfram Burgard. Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In *International Symposium on Experimental Robotics (ISER)*, 2016.

[114] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

[115] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[116] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5):1239–1285, 2020.

[117] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*, 2018.

[118] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 561–577. Springer, 2020.

[119] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. *arXiv preprint arXiv:2105.05633*, 2021.

[120] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[121] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[122] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[123] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[124] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021.

[125] Qiran Jia and Hai Shu. Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation. *arXiv preprint arXiv:2109.12271*, 2021.

[126] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7077–7087, 2021.

[127] Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie. Fastformer: Additive attention can be all you need. *arXiv preprint arXiv:2108.09084*, 2021.

[128] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[129] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[130] Qi Song, Jie Li, Chenghong Li, Hao Guo, and Rui Huang. Fully attentional network for semantic segmentation. *arXiv preprint arXiv:2112.04108*, 2021.

[131] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.

[132] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *arXiv preprint arXiv:2106.09681*, 2021.

[133] Yajie Xing, Jingbo Wang, and Gang Zeng. Malleable 2.5 d convolution: Learning receptive fields along the depth-axis for rgb-d scene parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 555–571. Springer, 2020.

[134] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. Spatial information guided convolution for real-time rgbd semantic segmentation. *IEEE Transactions on Image Processing*, 30:2313–2324, 2021.

[135] Weiyue Wang and Ulrich Neumann. Depth-aware cnn for rgb-d segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018.

[136] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531. IEEE, 2021.

[137] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3029–3037, 2017.

[138] Fahimeh Fooladgar and Shohreh Kasaei. Multi-modal attention-based fusion model for semantic segmentation of rgb-depth images. *arXiv preprint arXiv:1912.11691*, 2019.

[139] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[140] Eduardo Romera, José M. Álvarez, Luis M. Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272, 2018.

[141] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Gläser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2021.

[142] Apoorva Raghunandan, Mohana, Pakala Raghav, and H. V. Ravish Aradhya. Object detection algorithms for video surveillance applications. In *2018 International Conference on Communication and Signal Processing (ICCSP)*, pages 0563–0568, 2018.

[143] Yalong Pi, Nipun D. Nath, and Amir H. Behzadan. Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Adv. Eng. Informatics*, 43:101009, 2020.

[144] Kechen Song, Yingying Zhao, Liming Huang, Yunhui Yan, and Qinggang Meng. Rgb-t image analysis technology and application: A survey. *Eng. Appl. Artif. Intell.*, 120:105919, 2023.

[145] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Multispectral pedestrian detection via simultaneous detection and segmentation. In *British Machine Vision Conference*, 2018.

[146] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022.

[147] Fang Qingyun and Wang Zhaokui. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, 130:108786, 2022.

[148] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 403–411, June 2023.

[149] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021.

[150] Heng Zhang, Élisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. *2020 IEEE International Conference on Image Processing (ICIP)*, pages 276–280, 2020.

[151] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3489–3497, 2021.

[152] Yi Yu and Fei peng Da. Phase-shifting coder: Predicting accurate orientation in oriented object detection. *ArXiv*, abs/2211.06368, 2022.

[153] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *The European Symposium on Artificial Neural Networks*, 2016.

[154] Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch. Fully convolutional region proposal networks for multispectral person detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 243–250, 2017.

[155] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 139–158. Springer, 2022.

[156] Yang Zheng, Izzat H. Izzat, and Shahrzad Ziaee. Gfd-ssd: Gated fusion double ssd for multispectral pedestrian detection. *ArXiv*, abs/1903.06999, 2019.

[157] Heng ZHANG, Élisa Fromont, Sébastien Lefèvre, Bruno Avignon, and Universite de Rennes. Guided attentive feature fusion for multispectral pedestrian detection. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 72–80, 2021.

[158] Zijia An, Chunlei Liu, and Yuqi Han. Effectiveness guided cross-modal information sharing for aligned rgb-t object detection. *IEEE Signal Processing Letters*, 29:2562–2566, 2022.

[159] Xiaoxiao Yang, Yeqian Qiang, Huijie Zhu, Chunxiang Wang, and Ming Yang. Baanet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection. *2022 International Conference on Robotics and Automation (ICRA)*, pages 2920–2926, 2021.

[160] Chien-Yao Wang, Hong-Yuan Mark Liao, I-Hau Yeh, Yueh-Hua Wu, Ping-Yang Chen, and Jun-Wei Hsieh. Cspnet: A new backbone that can enhance learning capability of cnn. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1571–1580, 2019.

[161] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[162] Glenn Jocher. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements. `https://github.com/ultralytics/yolov5`, October 2020.

[163] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

[164] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 548–558, 2021.

[165] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *NeurIPS*, 2018.

[166] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. SPIGAN: Privileged adversarial learning from simulation. In *International Conference on Learning Representations*, 2019.

[167] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *ArXiv*, abs/1612.02649, 2016.

[168] Tuan-Hung Vu, Himalaya Jain, Max Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2512–2521, 2019.

[169] Rui Li, Wenming Cao, Qianfen Jiao, Si Wu, and Hau-San Wong. Simplified unsupervised image translation for semantic segmentation adaptation. *Pattern Recognition*, 105:107343, 2020.

[170] Dayan Guan, Jiaxing Huang, Shijian Lu, and Aoran Xiao. Scale variance minimization for unsupervised domain adaptation in image segmentation. *Pattern Recognition*, 112:107764, 2021.

[171] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[172] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, pages 1–15, 2021.

[173] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5981–5990, 2019.

[174] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021.

[175] Yajie Xing, Jingbo Wang, and Gang Zeng. Malleable 2.5d convolution: Learning receptive fields along the depth-axis for rgb-d scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[176] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Groß. Efficient rgb-d semantic segmentation for indoor scene analysis. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531, 2021.

[177] Hao Zhou, Lu Qi, Hai Huang, Xu Yang, Zhaoliang Wan, and Xianglong Wen. Canet: Co-attention network for rgb-d semantic segmentation. *Pattern Recogn.*, 124:108468, apr 2022.

[178] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016.

[179] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

[180] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1597–1607, 2020.

[181] Pengfei Zhu, Zhilin Zhu, Yu Wang, Jinglin Zhang, and Shuai Zhao. Multi-granularity episodic contrastive learning for few-shot learning. *Pattern Recognition*, 131:108820, November 2022.

[182] Jiabin Liu, Zhiquan Qi, Bo Wang, YingJie Tian, and Yong Shi. Self-llp: Self-supervised learning from label proportions with self-ensemble. *Pattern Recognition*, 129:108767, 2022.

[183] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[184] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *AAAI*, 2019.

[185] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[186] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, 2017.

[187] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021.

[188] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[189] Paolo Testolina, Francesco Barbato, Umberto Michieli, Marco Giordani, Pietro Zanuttigh, and Michele Zorzi. Selma: Semantic large-scale multimodal acquisitions in variable weather, daytime and viewpoints. *arXiv preprint arXiv:2204.09788*, 2022.

[190] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[191] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[192] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[193] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12409–12419, 2021.

[194] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[195] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *ArXiv*, abs/2203.04838, 2022.

[196] L. Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[197] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[198] Fan Liu, Delong Chen, Zhan-Rong Guan, Xiaocong Zhou, Jiale Zhu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *ArXiv*, abs/2306.11029, 2023.

[199] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Deep learning pour la fusion multimodale d'images : application à l'analyse de scènes extérieures dans des conditions difficiles

Mots clés: Fusion multimodale, l'apprentissage en profondeur, Analyse de scènes extérieur, Segmentation sémantique, Détection d'objet

Résumé: Les données visuelles multimodales peuvent fournir des informations différentes sur la même scène, améliorant ainsi la précision et la robustesse de l'analyse de scènes. Cette thèse se concentre principalement sur la façon d'utiliser efficacement les données visuelles multimodales telles que les images en couleur, les images infrarouges et les images de profondeur, et sur la façon de fusionner ces données visuelles pour une compréhension plus complète de l'environnement. Nous avons choisi la segmentation sémantique et la détection d'objets, deux tâches représentatives de la vision par ordinateur, pour évaluer et valider différentes méthodes de fusion de données visuelles multimodales. Ensuite, nous proposons un schéma de fusion RGB-D basé sur l'attention additive, considérant la carte de profondeur comme une modalité auxiliaire pour fournir des indices géométriques supplémentaires, et résolvant le coût élevé associé à l'auto-attention. Compte tenu de la complexité de la perception de scènes en conditions de faible luminosité, nous avons conçu un module de fusion croisée qui utilise l'attention de canal et spatiale pour explorer les informations complémentaires des paires d'images visible-infrarouge, améliorant ainsi la perception de l'environnement par le système. En fin, nous avons également abordé l'application des données visuelles multimodales dans l'adaptation de domaine non supervisée. Nous proposons d'utiliser des indices de profondeur pour guider le modèle à apprendre la représentation de caractéristiques invariables au domaine. Les nombreux résultats expérimentaux indiquent que les méthodes proposées surpassent les autres méthodes sur plusieurs bases de données multimodales disponibles publiquement et peuvent être étendues à différents types de modèles, démontrant ainsi davantage la robustesse et les capacités de généralisation de nos méthodes dans les tâches de perception de scènes en extérieur.

Title: Deep Multimodal Visual Data Fusion for Outdoor Scenes Analysis in Challenging Weather Conditions

Keywords: Multimodal fusion, Deep learning, Outdoor scene analysis, Semantic segmentation, Object detection

Abstract: Multi-modal visual data can provide different information about the same scene, thus enhancing the accuracy and robustness of scene analysis. This thesis mainly focuses on how to effectively utilize multi-modal visual data such as color images, infrared images, and depth images, and how to fuse these visual data for a more comprehensive understanding of the environment. Semantic segmentation and object detection, two representative computer vision tasks, were selected for investigating and verifying different multi-modal visual data fusion methods. Then, we propose an additive-attention-based RGB-D fusion scheme, considering the depth map as an auxiliary modality to provide additional geometric clues, and solving the high cost associated with self-attention. Considering the complexity of scene perception under low-light conditions, we designed a cross-fusion module that uses channel and spatial attention to explore the complementary information of visible-infrared image pairs, enhancing the system's perception of the environment. Additionally, we also researched the application of multi-modal visual data in unsupervised domain adaptation. We proposed to leverage depth cues to guide the model to learn domain-invariant feature representation. Extensive research results indicate that the proposed methods outperform others on multiple publicly available multi-modal datasets and can be extended to different types of models, which further demonstrating the robustness and generalization capabilities of our methods in outdoor scene perception tasks.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France