Multispectral object detection

Heng Zhang

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLÉ DOCTORALE Nº 601
Mathèmatique et Sciences et Technologies de l'Information et de la Communication
Spécialité : Informatique

Par
Heng ZHANG

MULTISPECTRAL OBJECT DETECTION

Thèse présentée et soutenue à Rennes le 14 decembre 2021
Unité de recherche : IRISA — UMR6074

Rapporteurs avant soutenance :
Pr. Vincent LEPETIT — Professor Univ Bordeaux, Researcher ENPC ParisTech
Pr. Tinne TUYTELAARS — Professor, ESAT KU Leuven

Composition du jury :

Président :	Dr. Patrick BOUTHEMY	Directeur de Recherche, INRIA
Examinateurs :	Dr. Patrick PEREZ	Senior Researcher, Valeo AI
	Dr. Jakob VERBEEK	Senior Researcher, Facebook
Dir. de thèse :	Pr. Elisa FROMONT	Professor, Université de Rennes 1
Co-dir. de thèse :	Pr. Sébastien LEFEVRE	Professor, Université de Bretagne-Sud

Invité :
Dr. Bruno AVIGNON — Manager, ATERMES company

Acknowledgements

First and foremost, I am very grateful to my supervisors Elisa Fromont and Sébastien Lefevre, for all your support and encouragement during my three-year PhD life. Despite your busy schedules, I sincerely appreciate your continued dedication to our weekly meetings, where you taught me how to think like a researcher, to formulate a research problem, and to present a research work.

I would especially like to express my thanks of gratitude to my industrial supervisor, Bruno Avignon. Thank you for your great support that makes this thesis possible. I am honoured to work with your R&D team, and I really appreciate your dedication to your work and care for your colleagues, that motivates me to always strive to make the project better

I would like to thank all my friendly, lovely and extremely smart colleagues in the INRIA laboratory, and I sincerely wish you more success in your future careers. (PS: I would be even more grateful if you consider citing my papers more XD.) Special thanks to the abacus-12 node of the IGRIDA server, where most of the methods proposed in this thesis are experimented.

Finally, my gratitude goes to my parents and my girlfriend Yutong YAN. Your company has been invaluable, and your love makes everything possible ♡.

Absorb what is useful, reject what is useless,
add what is essentially your own.

BRUCE LEE

Résumé étendu

Contexte et motivation

Les derniers développements technologiques en vision par ordinateur ont considérablement amélioré la capacité d'analyse automatique d'une scène, permettant ainsi de disposer de systèmes de vision *intelligents*. Par exemple, les méthodes de détection d'objets et de segmentation sémantique ont été appliquées à des systèmes de conduite autonome et de surveillance automatique. De plus, la fusion multi-capteurs a permis d'améliorer la fiabilité de reconnaissance dans différentes situations de la vie courante. Dans ce contexte, notre motivation est double. D'une part, nous souhaitons nous appuyer sur ces avancées récentes de la communauté scientifique pour construire des produits industriels basés sur l'analyse automatique de scène; d'autre part, la mise en oeuvre de ces solutions dans un contexte industriel permet d'identifier des problèmes pratiques auxquels les méthodes existantes sont confrontés, et de résoudre ces problèmes au travers de contributions originales que nous apportons à la communauté scientifique.

En particulier, nous nous concentrons dans cette thèse sur la tâche de détection d'objets. Celle-ci fournit une compréhension de base de la scène en identifiant tous les objets d'intérêt présents dans l'image. Concrètement, elle localise les objets par des boites englobantes et les associe à une classe particulière parmis un ensemble de classes prédéfinies. De nos jours, la plupart des modèles de détection d'objets basés sur l'apprentissage profond sont conçus pour améliorer la précision de la détection. Cependant, dans la plupart des cas, un modèle de détection plus précis nécessite davantage de paramètres et de calculs, car les réseaux de neurones plus profonds et plus larges sont à même d'extraire des caractéristiques plus discriminantes pour la tâche de reconnaissance. Cependant, du point de vue du déploiement industriel, les ressources de calcul des systèmes embarqués sont généralement limitées. Il est donc nécessaire de tenir compte de **l'efficacité de la détection**. Autrement dit, nous souhaitons obtenir une détection précise à partir de modèles compacts. Par conséquent, nous cherchons à améliorer la précision des modèles de détection existants sans introduire de calculs supplémentaires, et à réduire la complexité des modèles de détection pour les rendre plus adaptés aux systèmes embarqués.

Les systèmes de détection d'objets industriels nécessitent des détections précises dans diverses conditions, telles que l'obscurité, le contre-jour, la pluie, le brouillard, l'ombre, etc. Ces situations sont difficiles pour les systèmes utilisant uniquement des caméras visibles. Dans le but d'améliorer la fiabilité de la détection, les systèmes multispectraux peuvent introduire des caméras thermiques supplémentaires pour

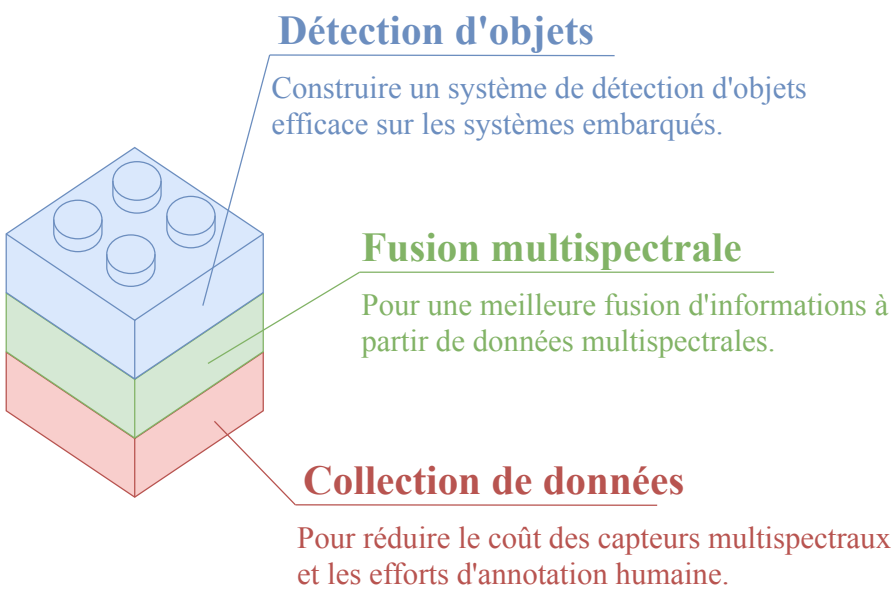


FIGURE 1 – Nos contributions pour la détection d'objets en imagerie multispectrale, présentées sous forme de "pièces de Lego".

combiner les informations provenant des deux modalités. Par rapport à la détection d'objets par imagerie monospectrale, la détection d'objets par imagerie multispectrale est un domaine de recherche beaucoup plus récent. Dans notre contexte, nous sommes particulièrement intéressés par **l'optimisation de la fusion d'informations multispectrales**. Par exemple, nous cherchons à intégrer efficacement la modalité thermique supplémentaire dans le cadre mature de détection d'objets monospectrale (RVB), afin de tirer pleinement profit de la complémentarité des deux modalités. En outre, notre démarche étant motivée par la résolution de problématiques opérationnelles, nous avons également souhaité réduire les coûts logiciels (par exemple, simplifier l'architecture du réseau profond), les coûts matériels (par exemple, réduire les besoins en caméras thermiques) et les coûts d'étiquetage (par exemple, sélectionner des échantillons plus pertinents pour l'annotation). À notre connaissance, ces aspects n'ont pas encore été étudiés dans le cadre de la détection d'objets en imagerie multispectrale.

Approches proposées

La conception d'un système efficace de détection d'objets par imagerie multispectrale nous a amenés à nous intéresser à trois problématiques : la détection d'objets, la fusion d'informations multispectrales, et la collecte efficace de données. Comme illustré en Figure 1, ces trois axes sont complémentaires et peuvent être vus comme des "pièces de Lego" pour construire le système de vision souhaité. Nous détaillons les contributions que nous avons apportées pour chacun de ces axes.

Dans le cadre de la problématique générale de détection d'objets, nous avons optimisé à la fois la précision et la vitesse des modèles profonds de détection. Pour cela, nous avons considéré deux stratégies alternatives. La première, nommée **Mutual Guidance**, attribue des étiquettes pour la tâche de classification en fonction de la qualité de prédiction sur la tâche de localisation et vice versa. Cette stratégie fournit non seulement une correspondance adaptative entre les ancres et les objets, mais elle atténue également le problème de non-alignement de prédiction entre les tâches de localisation et de classification. Notre deuxième contribution consiste à introduire une nouvelle méthode de compression de modèle nommée **PDF-Distil**, qui tire parti des désaccords de prédiction enseignant-étudiant pour guider le transfert de connaissances dans un cadre de distillation de connaissances pour la tâche de détection d'objets. En incorporant les informations de niveau prédiction, l'imitation au niveau caractéristiques se concentre automatiquement sur les zones où le modèle étudiant fait des prédictions inexactes, réduisant ainsi considérablement la complexité de calcul des modèles de détection d'objets.

Ensuite, dans le domaine de la fusion d'informations multispectrales, nous avons tenté de traiter le problème d'incohérence de modalités lors de la fusion. Les informations des caméras visibles et thermiques sont complémentaires, mais la fusion multispectrale devient difficile lorsque les deux caméras fournissent des informations contradictoires. Trois solutions ont été proposées pour faire face à cette situation. Premièrement, nous avons proposé un nouveau réseau de fusion nommé **Cyclic Fuse-and-Refine**, qui affine consécutivement les caractéristiques monospectrales avec les caractéristiques multispectrales fusionnées pendant le processus de fusion. Cette architecture de réseau réduit la différence entre les caractéristiques visibles et thermiques, et améliore ainsi la qualité globale des caractéristiques. Nous proposons ensuite un deuxième module de fusion multispectrale nommé **Progressive Spectral Fuse**, où les caractéristiques multispectrales sont progressivement fusionnées à travers plusieurs niveaux de convolution. La troisième contribution est pour sa part basée sur le mécanisme d'attention supervisée. Lorsque celui-ci est appliqué au cas de la fusion multispectrale, nous pouvons espérer que le réseau sélectionne activement la modalité avec une qualité de caractéristiques supérieure. Cependant, nous avons montré que le manque de supervision nuisait à la fusion basée sur l'attention, et nous avons proposé le modèle **Guided Attentive Feature Fusion** pour guider explicitement ce processus de fusion. Sans besoin d'annotations supplémentaires, notre méthode réalise une fusion entièrement adaptative des caractéristiques visibles et thermiques.

Enfin, afin d'assurer une collecte de données efficace, nous avons souhaité appliqué les mécanismes d'apprentissage actif et de distillation des connaissances au contexte de l'analyse de scène multispectrale. Nous avons étudié la complémentarité entre caméras multispectrales pour la **sélection active** de paires d'images multispectrales à annoter. Contrairement aux précédentes méthodes d'apprentissage actif où seules les images RVB sont prises en compte, nous nous intéressons plutôt à la différence de prédiction entre deux capteurs. De plus, afin de réduire le coût

matériel des systèmes d'analyse de scènes multispectrales, nous proposons un nouveau cadre de distillation de connaissances nommé **Modality Distillation**, qui distille la connaissance d'un réseau à deux branches à haute résolution thermique vers un seul à faible résolution thermique. Le modèle distillé est capable d'effectuer une prédiction précise sur les caméras thermiques à basse résolution, et affiche une complexité de calcul similaire à celle des modèles RVB uniquement.

Liste des publications

Les travaux présentés dans cette thèse ont fait l'objet des publications suivantes :

"Multispectral Fusion for object detection with Cyclic Fuse-and-Refine blocks" in *27th International Conference on Image Processing (ICIP2020)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

"Localize to Classify and Classify to Localize: Mutual Guidance in Object Detection", in *15th Asian Conference on Computer Vision (ACCV2020)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

"Guided Attentive Feature Fusion for Multispectral Pedestrian Detection", in *Winter Conference on Applications of Computer Vision (WACV2021)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

"Deep Active Learning from Multispectral Data Through Cross-Modality Prediction Inconsistency" in *28th International Conference on Image Processing (ICIP2021)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

"Low Cost Multispectral Scene Analysis with Modality Distillation", in *Winter Conference on Applications of Computer Vision (WACV2022)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

"PDF-Distil: including Prediction Disagreements in Feature-based Distillation for object detection", in *32nd British Machine Vision Conference (BMVC2021)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

Summary

Chapter 1

Introduction

Contents

The latest developments of computer vision technologies have greatly improved the ability of analysing scene and empowered various intelligent vision systems. For example, object detection and semantic segmentation methods have been applied to autonomous driving and automatic surveillance systems (Caesar et al., 2020; Cordts et al., 2016; Geiger et al., 2013; Leal-Taixé et al., 2015). In addition, multi-sensor fusion technology has improved the recognition reliability under different challenging situations. On the one hand, we are motivated to take advantage of these advances in the research community, to build industrial vision-based products; on the other hand, with the experience of product development, we are capable to identify practical problems from existing methods, and solve these problems by proposing our own contributions back to the research community. From our perspective, our research journey based on solving actual needs, bridges the research and industrial world, and proves that the two can promote each other.

1.1 Context and motivations

Recent progress on deep learning has achieved great successes on various computer vision tasks. In this thesis, we focus on the task of object detection in images or independent video frames. As shown in Figure 1.1, object detection aims at localizing objects through bounding boxes and assigning each of them to a predefined class. Object detection provides a basic understanding of the scene by identifying all existing objects of interest in the image. Object detection methods can be generally divided into non learning-based (Dalal & Triggs, 2005; Lowe, 1999; Viola & Jones, 2001) and learning-based (Bochkovskiy et al., 2020; Cai & Vasconcelos, 2018; Lin, Goyal, et al., 2017; W. Liu et al., 2016; Ren et al., 2016; Tian et al., 2019) methods. The key difference between the two is that for the former methods, features are usually manually defined; while for the latter methods, features are automatically learned via gradient descent. Today, the research field is almost entirely focused

FIGURE 1.1 – Example of prediction results for the object detection task from (H. Zhang, Fromont, Lefèvre, et al., 2020).

on the learning-based object detection methods, which are usually based on Convolutional Neural Networks (CNN). Although object detection is a long-established field, the research on this task is still active and the precision of object detection models has continued to improve rapidly in recent years, as shown in Figure 1.2.

In most of the cases, a more accurate detection model requires more parameters and calculations, since deeper and wider neural networks are capable to extract more discriminate features for recognition. However, from the industrial deployment standpoint, the computational resources of widely-used low-power embedded devices are generally limited compared to normal GPU servers. Indeed, most of the deep learning-based object detection models are designed and evaluated on GPU devices, with the pursuit of the precision of detection. We are, instead, more concerned about **the efficiency of the detection**, i.e., we prefer precise detection from compact models. Therefore, we are interested in improving the precision of existing detection models without introducing additional calculations, and reducing detection models' complexity to make them more fit to edge devices. We believe that the improvements on these aspects are as crucial as precision, for the application of deep learning-based detection models on industrial products.

Although today's object detection models achieve excellent performance on public datasets, it still remains unknown whether they could maintain high reliability when applied to real-life situations. Industrial object detection systems require accurate detection performance under various conditions, such as darkness, backlight,

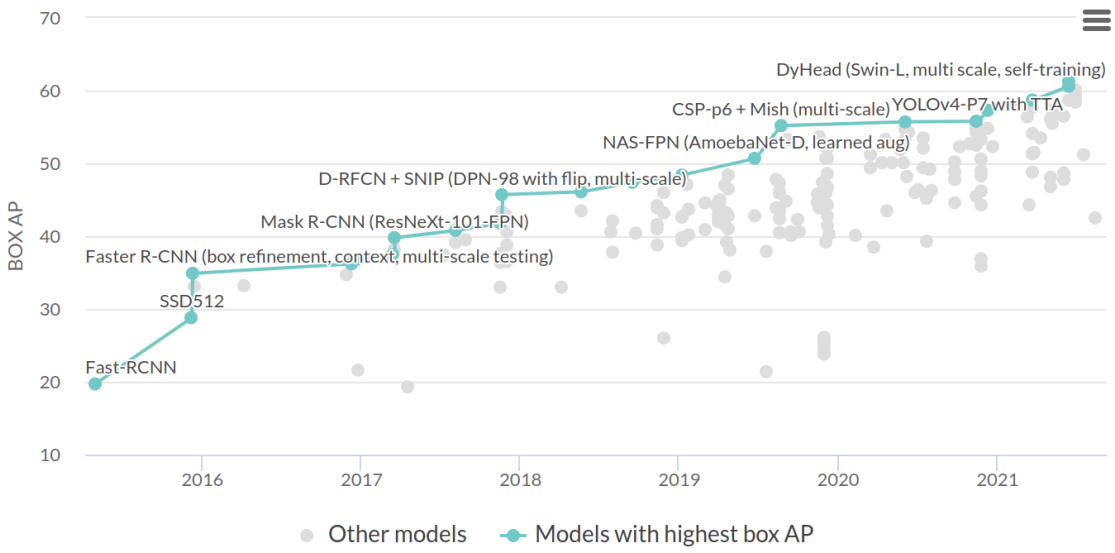


FIGURE 1.2 – Evolution of object detection precision on COCO dataset since 2015. The chart is taken from paperswithcode.com.

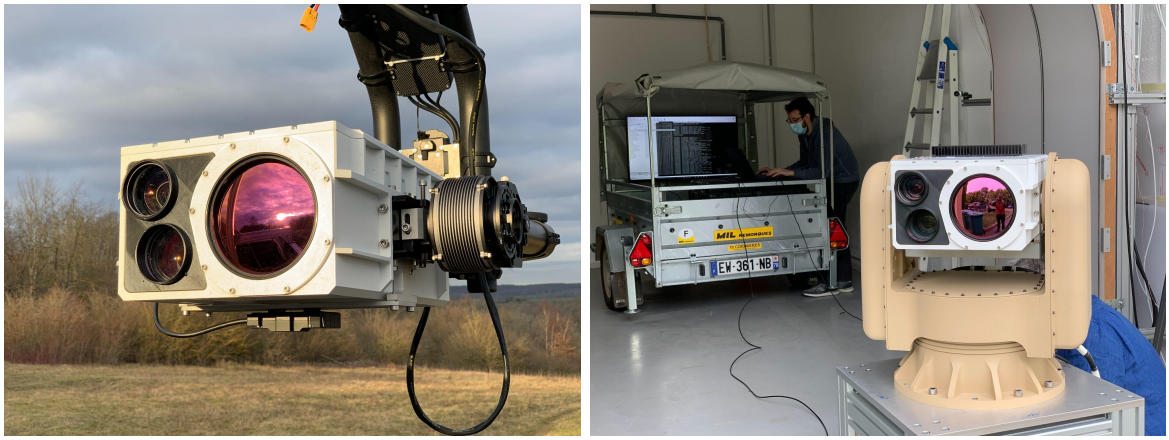


FIGURE 1.3 – The multispectral surveillance system named SIAMM from ATERMES. We show the system on moving (left) to collect training images, and the system on base (right) for actual deployment.

rain, fog, shadow, etc. These situations are challenging for systems using only standard visible cameras. With the purpose of improving the detection reliability, multispectral systems may introduce **additional thermal cameras** to combine the information coming from both modalities. Figure 1.3 shows one of the multispectral video surveillance products from ATERMES, the company which financed this thesis's research. On this system, the visible camera and the thermal camera are placed in parallel to provide multispectral image pairs with identical perception fields. Figure 1.4 demonstrates some real images captured by the system. As is shown, the conventional visible cameras provide precise visual details (such as colour and texture) in a well-lit environment, while the additional thermal cameras produce temperature maps of the scene according to objects' infrared radiation, which is particularly useful for object detection at nighttime or in the shade. In fact, the contributions from both cameras are complementary, which means that when one camera is

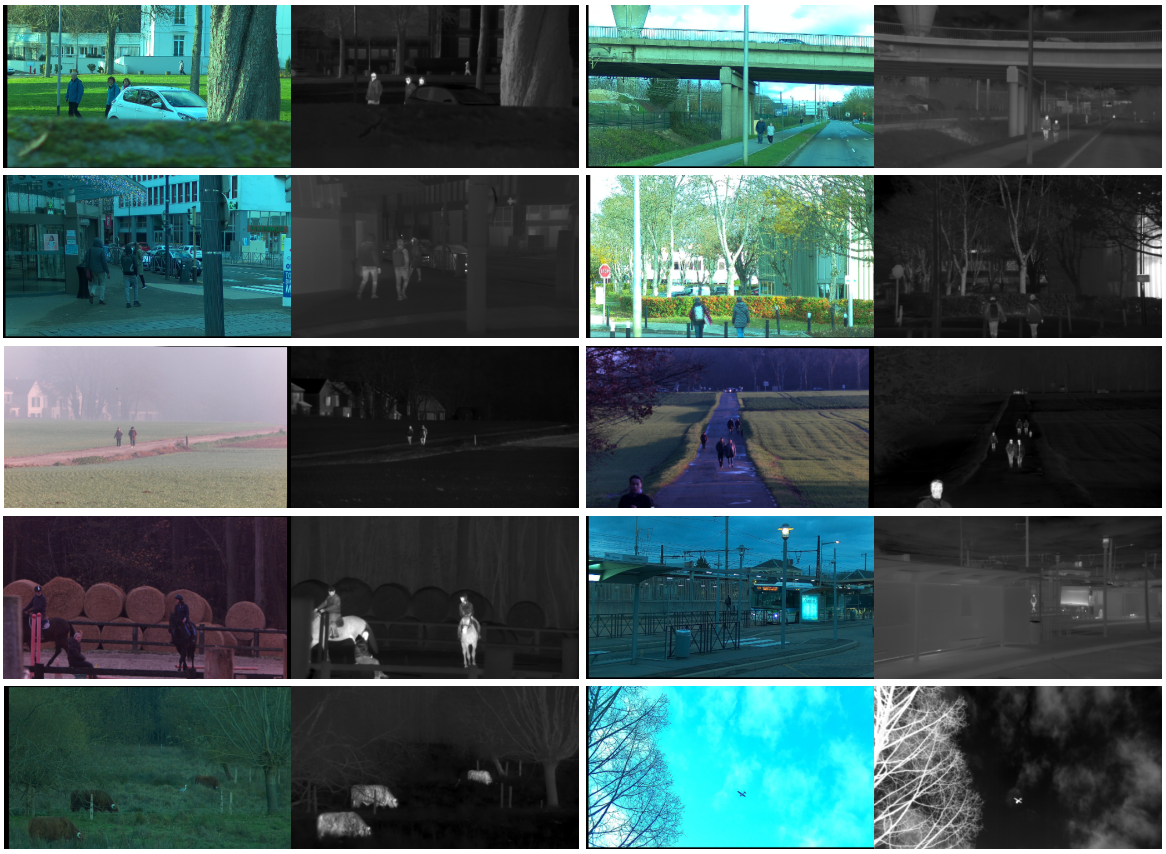


FIGURE 1.4 – Example of multispectral images pairs captured by SIAMM system from ATERMES.

in suboptimal imaging conditions, the other one can produce higher quality images to ensure reliable detection performance.

Compared with general object detection, multispectral object detection is a much younger research field. For instance, the first large-scale multispectral pedestrian detection dataset was published in 2015[1]. Since then, various multispectral object detection methods have been proposed and evaluated on this dataset. Today, the major research focus is on the **multispectral information fusion** process, i.e, how to efficiently integrate the additional thermal modality into the mature mono-modal (RGB) object detection framework, such that the complementarity of the two sensors can be exploited to the maximum extent possible. Specifically, the multispectral fusion has been tested in different ways and at different stages of the object detection pipeline. In general, the research on multispectral object detection is still in its infancy, thus it is very important to challenge the inherent thinking of the predecessors or even overthrow the established conclusions. Moreover, since our research is largely based on solving actual needs, we are also interested in reducing software costs (e.g., simplifying network architecture), hardware costs (e.g., reducing thermal camera requirements) and labelling costs (e.g., selecting valuable samples for annotation). To the best of our knowledge, these aspects have not been studied under the context of multispectral object detection yet.

[1]https://sites.google.com/site/pedestrianbenchmark/

1.2 Thesis outline

This thesis is organized as follows:

In Chapter 2, we present the basic concepts about general object detection (Section 2.1), multispectral fusion (Section 2.2), knowledge distillation (Section 2.3) and active learning (Section 2.4). These concepts will be used in the rest of this thesis. For each research topic, we discuss the limitations that we found when integrating them into actual industrial developments. To cope with the found limitations, we introduce our own contributions for each of the aforementioned domains in the following chapters of this thesis. Specifically, we assort our works into three "Lego bricks", corresponding to detection, fusion, and data.

In Chapter 3, we improve the precision of existing object detection models by introducing a novel **label assignment strategy** (Section 3.2). This work was presented at ACCV 2020 (H. Zhang, Fromont, Lefèvre, et al., 2020) and inspired two ICCV 2021 papers (Feng et al., 2021; Gao et al., 2021). Then, we manage to reduce the computational complexity of existing models by introducing a **detection-specific knowledge distillation framework** (Section 3.3). This framework works better than previous detection distillation methods and is accepted to BMVC 2021.

In Chapter 4, three different fusion strategies are presented to take advantage of the complementarity of multispectral features. Specifically, the first method designs a **cascaded network architecture** to refine the visible and the thermal features with the fused features (Section 4.1), the second method **progressively** fuses multispectral features throughout several convolution levels (Section 4.2), and the third method adapts the **attention mechanism** to the fusion module and introduces a specific supervision for training (Section 4.4). The first and the third works were presented at ICIP 2020 (H. Zhang, Fromont, Lefevre, et al., 2020) and WACV 2021 (H. Zhang et al., 2021a).

In Chapter 5, we propose some practical solutions for industrial purpose, such as reducing the labour and manufacture costs of constructing intelligent multispectral scene analysis systems. To be more specific, an **active learning** strategy for multispectral scene analysis (Section 5.1) is proposed based on the cross-modality prediction inconsistency. This strategy is evaluated on various multispectral vision tasks and was presented at ICIP 2021 (H. Zhang et al., 2021b). Moreover, a novel **knowledge distillation** framework (Section 5.2) is proposed under the context of multispectral scene analysis, where both the hardware and the software costs of a multispectral system are reduced. This work will be presented at WACV 2022 (H. Zhang et al., 2022) and will be patented.

Finally, to show the broad impact of our work, we also extend our methodologies to the field of **remote sensing**. We conclude this thesis by summarizing all our contributions and discussing potential **future works** in Chapter 6.

Chapter 2

Deep learning background

Contents

In this chapter, we introduce the necessary background and related work to understand the main concepts used in this thesis[1]. Concretely, we first summarize some representative works about general and multispectral object detection, knowledge distillation and active learning. For each field, we briefly review its research progress and point out the limitations of existing methods. These limitations motivated us to explore the essential causes of detection failures and propose corresponding solutions. Then, we present the public datasets used in this thesis.

2.1 General object detection

Object detection is one of the fundamental tasks in computer vision. The objective is to localize objects through bounding-boxes and assign each of them to a predefined class. Nowadays, deep learning-based (DL-based) methods largely dominate this research field. Modern DL-based object detection networks consist of three subnetworks: the backbone, the neck and the head. **Backbone networks** are used to extract features from the input images. They are often image classification networks, such as VGG series (Ding et al., 2021; Simonyan & Zisserman, 2015), ResNet series (He et al., 2016; Xie et al., 2017), MobileNet series (A. Howard et al., 2019; A. G. Howard et al., 2017; Sandler et al., 2018) and ShuffleNet series (N. Ma et al., 2018; X. Zhang et al., 2018). **Neck networks** realize multiscale object detection by fusing features at different scales. FPN (Lin, Dollár, et al., 2017) and PAFPN (S. Liu et al., 2018) are nowadays the most commonly adopted neck networks. **Head networks**

[1]We assume that the reader is familiar with basic notions about deep learning and computer vision, where more information can be found in (LeCun et al., 2015)

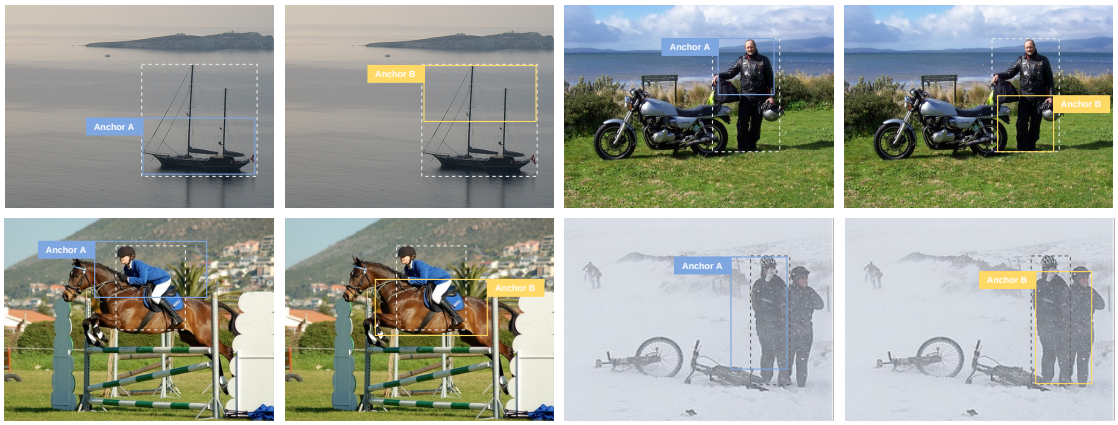


FIGURE 2.1 – Anchors A and anchors B have the same IoU with ground truth box, but different visual semantic information. The ground truth in each image is marked as dotted-line boxes. Better viewed in colour.

handle instance classification and bounding-box regression. They can be roughly divided into two types: two-stage and single-stage detectors. Two-stage detectors (Cai & Vasconcelos, 2018; Pang et al., 2019; Ren et al., 2016) firstly generate a variety of regions of interest, then refine and classify each region candidate separately; Single-stage detectors (Lin, Goyal, et al., 2017; W. Liu et al., 2016; Redmon et al., 2016) directly localize and classify all existing objects on the image. Another criterion divides head networks into anchor-based and anchor-free detectors. Anchor-based detectors (Lin, Goyal, et al., 2017; W. Liu et al., 2016; Ren et al., 2016) resort to numerous predefined anchor boxes (explained in the next paragraph) to handle objects' scale and shape variations; Anchor-free detectors directly predict objects' key-points (Duan et al., 2019; Law & Deng, 2018; X. Zhou, Zhuo, et al., 2019) or center-points (Tian et al., 2019; X. Zhou, Wang, et al., 2019; C. Zhu et al., 2019), without the help of anchor boxes. Although there exist various detection methods, according to our experiments, anchor-based single-stage method is currently the best choice for real-time object detection on embedded devices (the context of projects at ATERMES), because of its excellent accuracy/speed trade-off.

In order to train an anchor-based single-stage detection head, the matching between the predefined anchor boxes and the ground truth boxes is an inevitable step. To be more specific, anchors are predefined reference boxes of different sizes and aspect ratios uniformly stacked over the whole image. They help the network to handle objects' scale and shape variations by converting the object detection problem into an anchor-wise bounding-box regression and classification problem. Previous anchor-based object detectors resort to the Intersection-over-Union (IoU) between predefined anchor boxes and ground truth boxes (called IoU_{anchor} in the following) to assign the sample anchors to an object (positive anchors) or a background (negative anchors) category. These assigned anchors are then used to minimize the bounding-box regression and classification losses during training. This IoU_{anchor}-based anchor matching criterion is reasonable under the assumption that anchor boxes with high IoU_{anchor} are appropriate for the detection of the corresponding objects. However, in reality, the IoU_{anchor} is **insensitive to objects' content/context**, thus not "optimal" to be used, as such, for anchor matching.

In Figure 2.1, we show several examples where IoU_{anchor} does not well reflect the matching quality between anchors and objects: Anchors A and anchors B have exactly the same IoU_{anchor} but possess very different matching qualities. For example, on the first line of Figure 2.1, anchors A covers a more representative and informative part of the object than anchors B; On the second line, anchors B contains parts of a nearby object which hinders the prediction on the jockey/left person.

DL-based object detection involves two sub-tasks: instance localization and classification. Predictions for these two tasks tell us "where" and "what" objects are on the image, respectively. During the training phase, both tasks are jointly optimized by gradient descent, but the static anchor matching strategy does not explicitly benefit from the joint resolution of the two tasks, which may then yield to a **task-misalignment problem**, i.e., during the evaluation phase, the model might generate predictions with correct classification but imprecisely localized bounding-boxes as well as predictions with precise localization but wrong classification. Both predictions significantly reduce the overall detection quality.

To address the aforementioned two limitations of the existing static matching strategy, we question this use of IoU_{anchor} and propose, in Section 3.2, a new anchor matching criterion which is mutually guided by the prediction on the localization and the classification tasks: the predictions related to one task are used to dynamically assign sample anchors and improve the model on the other task, and vice versa.

2.2 Multispectral object detection

Video surveillance applications need to maintain high reliability under various conditions. However, situations such as insufficient illumination or adverse weather can be challenging for systems using only visible cameras, which is why multispectral systems introduce additional thermal cameras to provide supplementary information. In particular, visible cameras provide visual details on objects' colour and texture, while thermal cameras are sensitive to objects' temperature changes. The contributions of both cameras are complementary, and their combination can ensure reliable performance round-the-clock.

To effectively use the information from multispectral cameras, the main technical problem resides in the multispectral fusion process. According to the specific fusion stage, this fusion process can be divided into three categories: image-level fusion, feature-level fusion and decision-level fusion.

The first application of DL-based multispectral object detection is introduced in (Wagner et al., 2016). The authors compared the image-level and the feature-level fusion strategies, where the image-level fusion combines the multispectral information by directly concatenating thermal and RGB images, and the feature-level fusion applies a two-stream architecture (explained in the next paragraph and shown in Figure 2.2). Their conclusion was that **feature-level fusion produces superior**

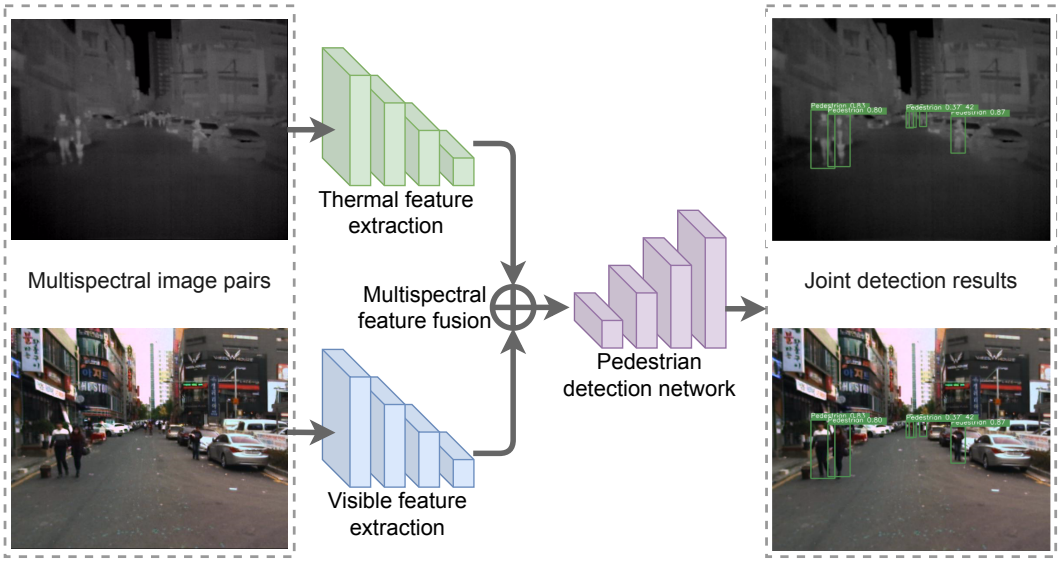


FIGURE 2.2 – Multispectral object detection via a two-stream convolutional neural network.

performance, whereas image-level fusion cannot even surpass traditional methods (such as Aggregated Channel Features (Dollár et al., 2014)). To the best of our knowledge, worldwide research on image-level fusion for multispectral object detection has been mostly interrupted since these findings. The research focus has then shifted to feature-level fusion.

Most feature-level fusion models adopt a **two-stream network architecture** (not to be confused with the two-stage detection network presented in Section 2.1). As illustrated in Figure 2.2, a typical two-stream object detection network consists of two separate spectra-specific feature extraction branches, a multispectral feature fusion module and an object detection network operating on the fused features. The model takes some aligned thermal-visible image pairs as input and outputs the joint detection results on each image pair. Various studies have been conducted based on this specific network architecture: Both (J. Liu et al., 2016) and (Konig et al., 2017) adapted Faster R-CNN (Ren et al., 2016) to a two-stream architecture for multispectral object detection. They compared different multispectral fusion stages and came to the conclusion that the fusion in the middle stage of the neural network outperforms the fusion in the early or late stages; Based on this, MSDS-RCNN (C. Li et al., 2018) adopted a two-stream middle-stage fusion architecture and combined the object detection and the semantic segmentation task to further improve the detection accuracy. CIAN (L. Zhang, Liu, Zhang, et al., 2019) applies a channel-level attention in the multispectral feature fusion stage to model the cross-modality interaction and weigh the visible and thermal features in the feature fusion stage; MBNet (K. Zhou et al., 2020) proposes the Differential Modality Aware Fusion (DMAF) module to alleviate the inconsistency between visible and thermal features and to facilitate the optimization process of a dual-modality network.

Apart from these approaches on image-level and feature-level fusion, multiple decision-level fusion methods were suggested: Both (Guan et al., 2019; C. Li et al.,

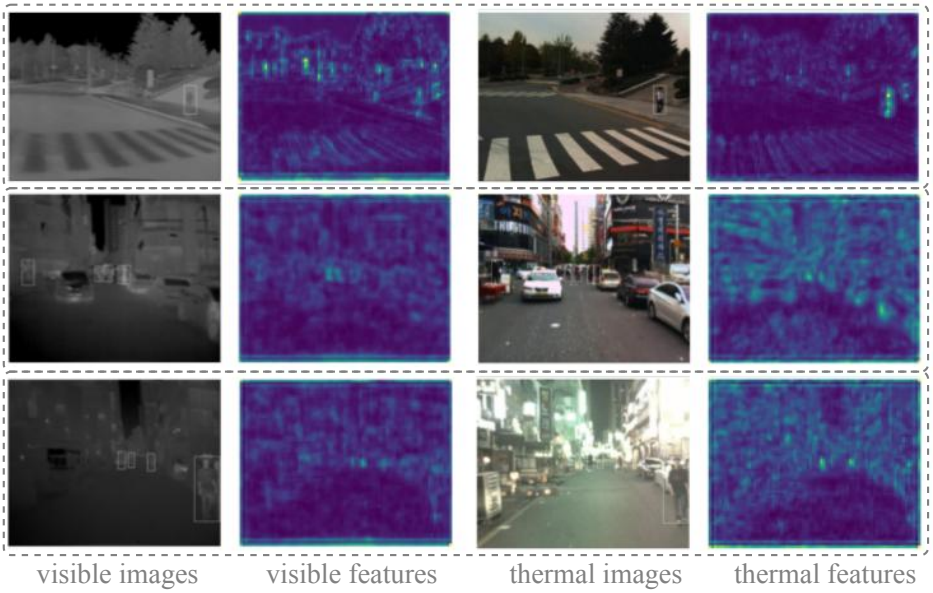


FIGURE 2.3 – Examples of multispectral image pairs and their corresponding features. It can be observed that the multispectral features are quite different.

2019) used illumination information as a clue to guide the fusion of predictions (decisions) from thermal/visible modalities. They train a separate network to estimate the illumination value from a given image pair, then (C. Li et al., 2019) uses the predicted illumination value to weigh the detection results from both the thermal and RGB images. (Guan et al., 2019) uses the illumination value to weigh the detection results from a day-illumination subnetwork and a night-illumination subnetwork. AR-CNN (L. Zhang, Zhu, et al., 2019) discussed a confidence-aware fusion mechanism, where the disagreement between visible and thermal predictions is used to re-weigh visible contributions, which could also be regarded as a decision-level fusion approach.

In our opinion, since thermal and visible cameras have different imaging characteristics under different conditions, the real difficulty of multispectral fusion resides in the cases where **the two cameras produce contradictory representations**. We demonstrate the existence of these contradictory representations via visualizing the input multispectral image pairs and their corresponding extracted features in Figure 2.3. It can be observed that, even though the multispectral image pairs are well aligned, the extracted thermal and visible features are quite different, and these representation disagreements may lead to uncertain and error-prone multispectral fusion results.

To better deal with these contradictory representations, we propose three solutions: In Section 4.1 and 4.2, we present two novel network architectures to progressively decrease the difference between thermal and visible features; In Section 4.4, we regard the multispectral fusion as a sub-task of the network optimization and introduce a specific guidance in the fusion module.

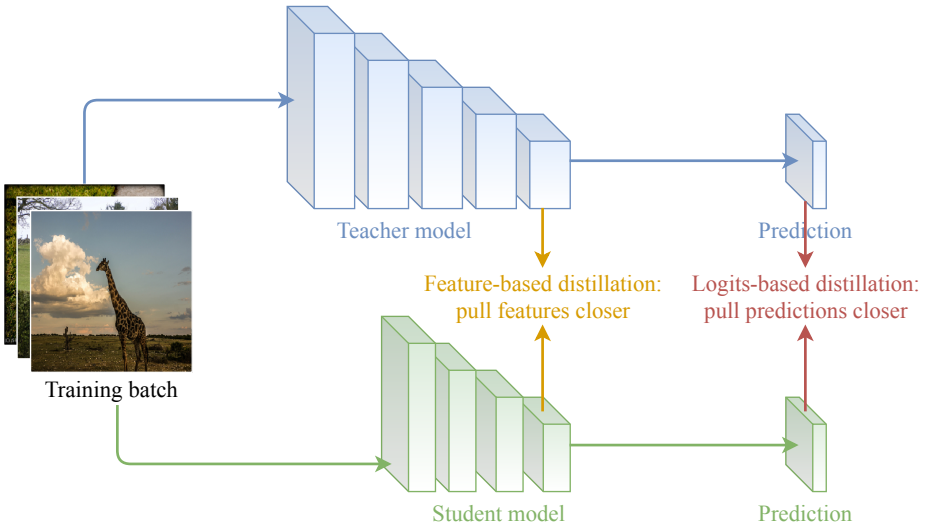
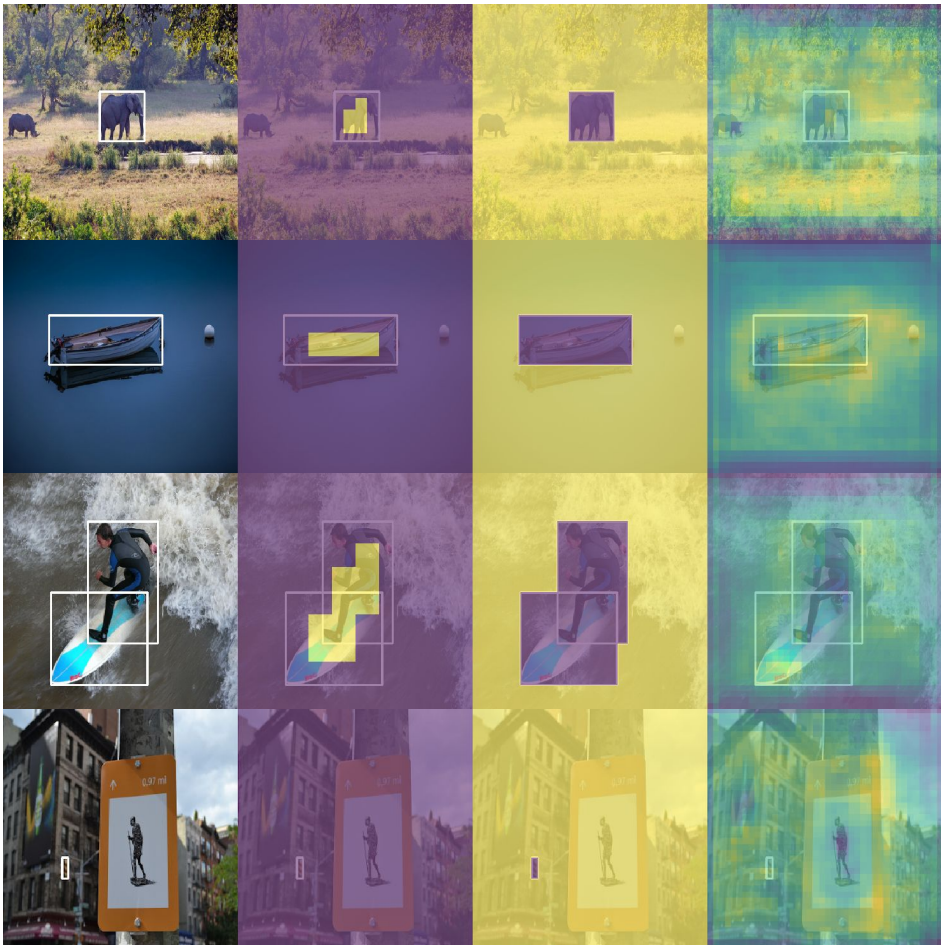


FIGURE 2.4 – Logits-based distillation and feature-based distillation are the two major knowledge transfer strategies in the literature.

2.3 Knowledge distillation

Despite their outstanding performance on different computer vision tasks, DL-based techniques still suffer from practical limitations that make them difficult to deploy at a large scale, especially when dealing with real-time scene analysis applications on embedded devices. This is due to the fact that state-of-the-art DL-based models often require enormous parameters and calculations, which make them both heavy to store and relatively slow at inference. Therefore, model compression techniques such as network pruning (Z. Liu et al., 2017), parameter quantification (Han et al., 2015) and knowledge distillation (Hinton et al., 2015) are suggested to reduce the storage cost and the computational complexity of deep models, while minimizing the performance degradation.

Specifically, knowledge distillation (KD) aims at compressing deep models by transferring the learned knowledge from a precise but cumbersome model to a compact one. A typical KD framework consists of three components: a complex teacher model, a compact student model and a knowledge transfer module. As illustrated in Figure 2.4, logits-based distillation and feature-based distillation are the two major knowledge transfer strategies. Logits-based methods (Hinton et al., 2015; Y. Zhang et al., 2018) assign the output probability from the teacher model as the (soft) target for the training of the student model. The motivation behind these logits-based methods is that, for a given input image, we expect the output prediction of the student model to be as similar as possible to that of the teacher model. Alternatively, feature-based methods (Romero et al., 2015; Zagoruyko & Komodakis, 2017) transfer high-level semantic information by making the student model mimic the internal representations (i.e. the features maps) of the teacher model. Deep features in neural networks carry rich semantic information, thereby providing better distillation guidance than the probability distributions. Leveraging this, feature-based methods are the most commonly adopted KD strategy for object detection (Guo et al., 2021; T. Wang et al., 2019; Y. Zhang et al., 2020).



(A) Input image (B) Fine-grained (C) Decoupled (D) Prime-aware

FIGURE 2.5 – Visualization of different sampling strategies for feature-based detection distillation. We plot from left to right: (A) input image with ground truth boxes, (B) Fine-grained (T. Wang et al., 2019), (C) Decoupled (Guo et al., 2021) and (D) Prime-aware (Y. Zhang et al., 2020).

However, when directly applying feature-based distillation to object detection models, the precision gap between the teacher model and the student model remains significant. As shown in Figure 2.5 (A), in the object detection task, the target objects normally only occupy a small part of the images. Therefore, the supervision of the feature distillation is often dominated by the abundant, less informative background. This foreground-background imbalance greatly reduces the efficiency of the knowledge transfer in feature-based distillation.

Several solutions shown in Figure 2.5 have been proposed to tackle this imbalance problem: Fine-grained (T. Wang et al., 2019) (B) suggested to only perform feature imitation on near object regions; Decoupled (Guo et al., 2021) (C) noticed that distillation on background regions reduces false positive detections, and thereby proposed to assign different weighting values for foreground features and for background features; Prime-aware (Y. Zhang et al., 2020) (D) realized an adaptive sample weighting by incorporating the uncertainty learning into the feature distillation. In their implementation, sample weighting is biased towards “easy” samples, where

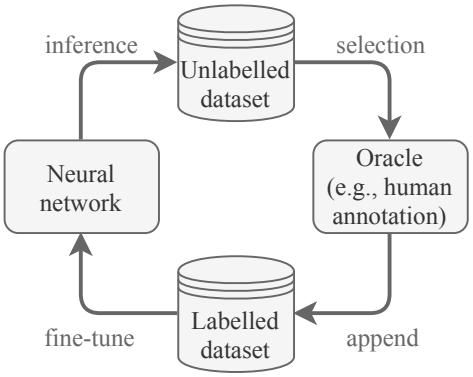


FIGURE 2.6 – Diagram for a typical active learning cycle.

most of which are actually background. We found that these methods are limited to feature-level operations, while discarding the initial motivation of KD, which is **minimizing the prediction difference** between the teacher and the student models. In Section 3.3 of this thesis, we will introduce an efficient sample weighting mechanism via combining feature-based with logits-based distillation, where the former is guided by the latter to more important areas on the feature maps.

2.4 Active learning

Labelled data are critical for today's supervised learning to realize a precise and reliable object detection. While there exist many large-scale benchmarks acquired by regular visible cameras, collecting labelled multispectral data is more expensive and time-consuming, e.g., acquiring well-aligned multispectral image pairs requires specific equipment, and few open datasets acquired with a similar equipment can be used as supplementary data when training multispectral models. Active Learning (AL), which aims to relieve human labelling efforts, is thus particularly appealing for our multispectral context. The AL protocol usually starts by pre-training a model on a small subset of the labelled dataset D_l. Then, several AL cycles are repeated. Figure 2.6 illustrates a typical AL cycle. The model inference is performed on the unlabelled dataset D_u to select the most informative samples (i.e., multispectral image pairs in our work). These selected samples are then sent to an external oracle for annotation and appended to the labelled dataset D_l, where the model is consequently fine-tuned on. The most important component of an AL cycle is the scoring function, which ranks the informativeness of unlabelled samples.

Most studies on deep AL in computer vision are based on mono-modal RGB images, including the most recent ones in deep AL for object detection (Aghdam et al., 2019; Brust et al., 2018; Kao et al., 2018; Roy et al., 2018) and semantic segmentation (Casanova et al., 2020; Siddiqui et al., 2020). Conversely to these existing works that score the informativeness of a single image, we suggest **relying on the complementarity of images from different cameras** for the adaptive selection of multispectral samples to be annotated. To the best of our knowledge, this is the first effort in deep AL within the context of multispectral scene analysis. More details about this novel AL approach will be given in Section 5.1.

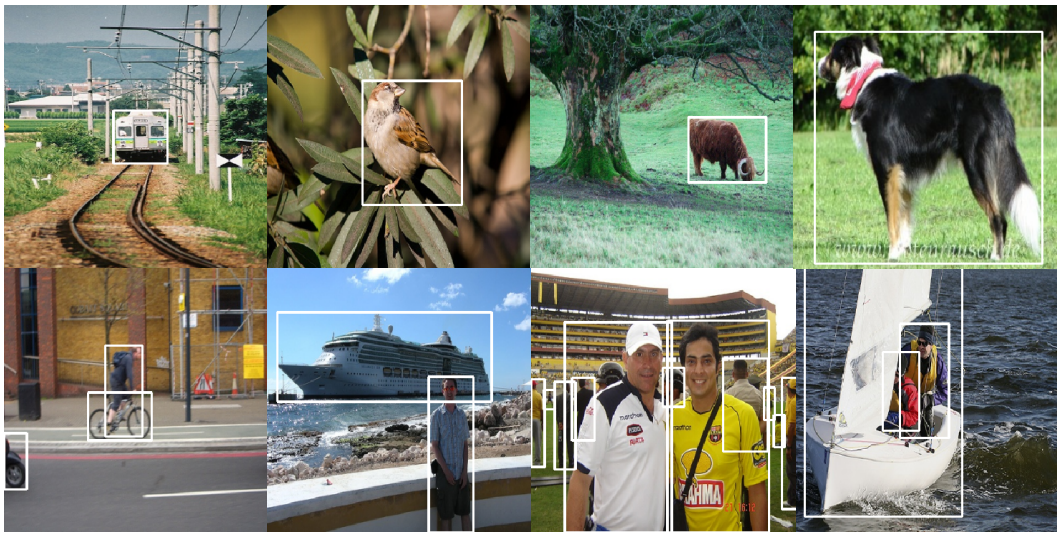


FIGURE 2.7 – Example of images from PASCAL VOC dataset.

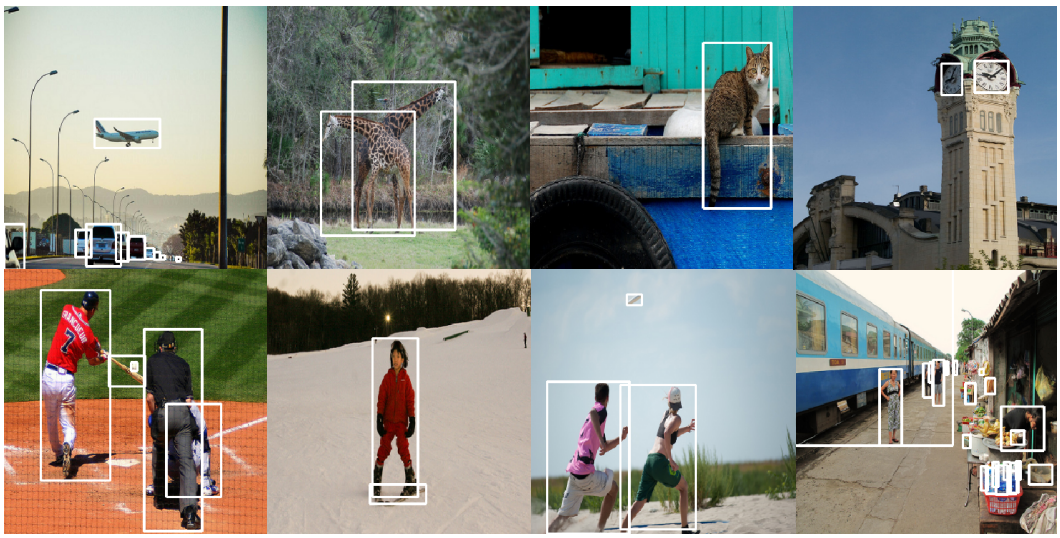


FIGURE 2.8 – Example of images from MS COCO dataset.

2.5 Datasets

2.5.1 General object detection datasets

Experiments for general object detection are conducted on two commonly-used annotated benchmark datasets: PASCAL VOC (Everingham et al., 2010) and MS COCO (Lin et al., 2014). Some examples of images from these two datasets and their annotated bounding-boxes are shown in Figure 2.7 and in Figure 2.8. PASCAL VOC dataset contains 20 object categories. Following the common practice, we utilize the combination of VOC2007 and VOC2012 trainval sets for training (16,551 images in total), and rely on the VOC2007 test for evaluation (4,952 images). MS COCO dataset contains 80 object categories. For fair comparisons with previous works, we use the train2017 set for training (117,244 images) and the val2017 set for evaluation (5,000 images).

FIGURE 2.9 – Example of multispectral image pairs from KAIST and FLIR datasets.

For both datasets, we adopt the (COCO-style) mean Average Precision (denoted as mAP) as the evaluation metric, which is defined as the average of AP scores across 10 Intersection-over-Union[2] (IoU) thresholds from 0.5 to 0.95. Moreover, we also report $AP50$, $AP75$, AP_s, AP_m and AP_l for comprehensive comparisons. $AP50$ and $AP75$ measure the average precision for a given IoU threshold (50% and 75%, respectively). The last three aim at evaluating detection precision on small ($area < 32^2 pixels$), medium ($32^2 pixels < area < 96^2 pixels$) and large ($area > 96^2 pixels$) objects respectively. Since the size of the objects greatly varies between MS COCO and PASCAL VOC, these size-dependent measures are ignored when experimenting only with PASCAL VOC dataset.

2.5.2 Multispectral object detection datasets

In order to evaluate the effectiveness of the proposed multispectral fusion methods, we conduct experiments on KAIST Multispectral Pedestrian Detection Dataset (Hwang et al., 2015) (denoted as KAIST dataset) and FLIR ADAS Dataset[3] (denoted as FLIR dataset). Some examples of multispectral image pairs from these two datasets are shown in Figure 2.9.

KAIST dataset focuses on the pedestrian detection task based on aligned multispectral image pairs. These image pairs are collected during daytime and nighttime.

[2]Intersection-over-Union (IoU) is a metric for measuring the distance between two bounding boxes, more information can be found in https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/.

[3]https://www.flir.com/oem/adas/adas-dataset-form/

This dataset contains 21,622 annotated image pairs for training, and 2,252 image pairs for testing. Due to some problematic annotations in the original dataset, several researchers (C. Li et al., 2018; J. Liu et al., 2016; L. Zhang, Zhu, et al., 2019) have proposed some improved annotations for training and evaluation. Following previous work, we adopt the Miss Rate (computed by averaging the miss rate on false positive per-image points sampled within the range of $\left[10^{-2}, 10^{0}\right]$, lower is better) under a "reasonable" setting (Dollar et al., 2011), i.e., only pedestrians taller than 50 pixels under no or partial occlusions are considered. In practice, we use the evaluation code provided by (C. Li et al., 2018)[4].

FLIR dataset is a recently released multispectral (multi-)object detection dataset. We only kept the 3 more frequent classes which are "bicycle", "car" and "person". Originally, it contains around 10k manually-annotated thermal images with their corresponding reference RGB images, collected during daytime and nighttime. We manually removed the misaligned visible-thermal image pairs and ended with 4,129 well-aligned image pairs for training and 1,013 image pairs for testing[5]. Models trained on this dataset are evaluated with the aforementioned mean Average Precision (mAP) metric.

2.5.3 Multispectral semantic segmentation dataset

MFNet dataset (Ha et al., 2017) targets the semantic segmentation of street scenes for the development of the advanced driver assistance systems (ADAS). The segmentation labels consist of eight classes: car, person, bike, curve, car stop, guardrail, colour cone and bump. The dataset provides 1,568 aligned multispectral image pairs in the training set, 392 pairs in the validation set and 393 pairs in the test set. Among each subset, half of the image pairs are taken during daytime, and the other half during nighttime. Visible and thermal images are again well aligned. Some examples of multispectral images and their ground truth segmentation masks from MFNet dataset are shown in Figure 2.10. To evaluate the segmentation accuracy, we report the class-wise Mean Accuracy, calculated by averaging the ratio between the number of true positive pixels and the sum of true positive and false negative pixels for each class.

[4]https://github.com/Li-Chengyang/MSDS-RCNN/tree/master/lib/datasets/KAISTdevkit-matlab-wrapper

[5]This "aligned" version of FLIR dataset can be downloaded at: https://drive.google.com/file/d/1xHDMGl6HJZwtarNWkEV3T4O9X4ZQYz2Y/view

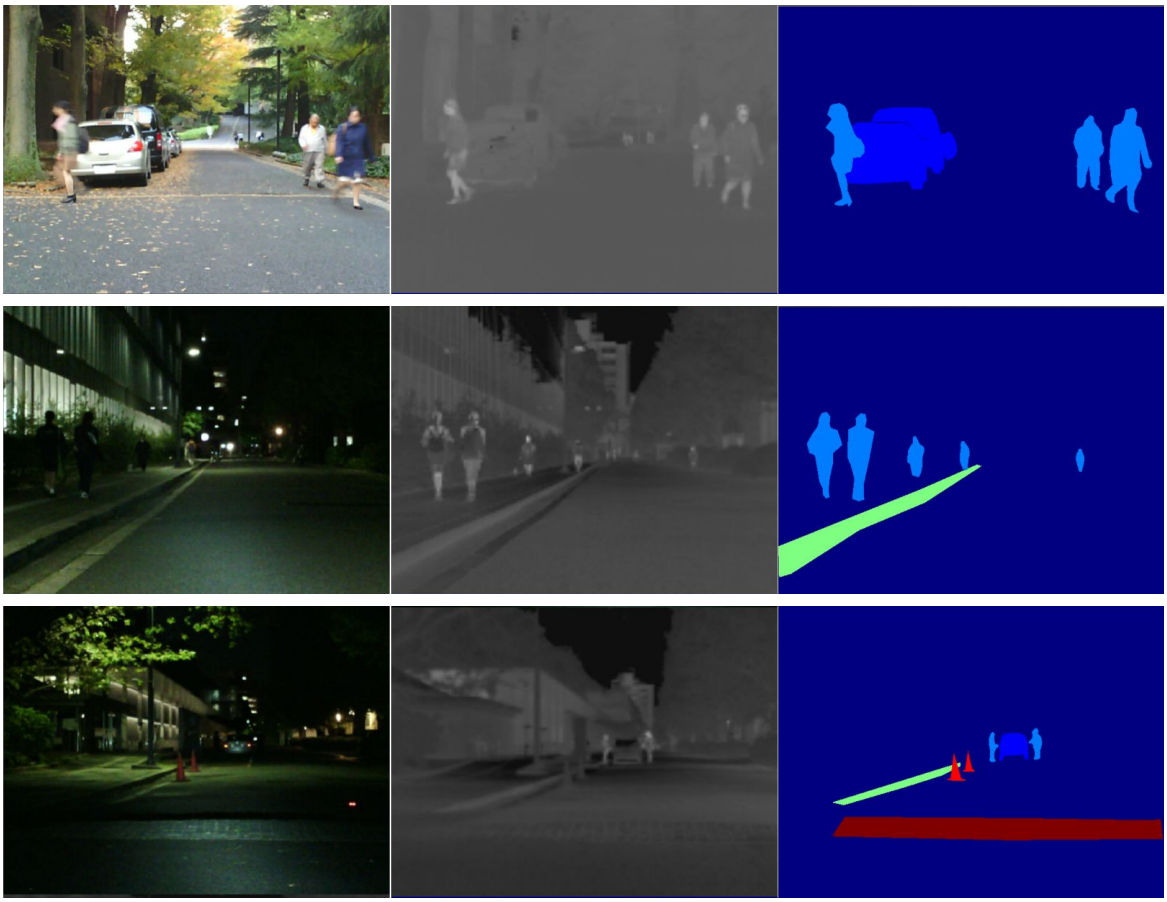
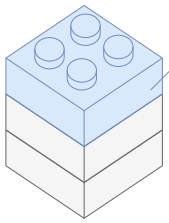


FIGURE 2.10 – Example of multispectral image pairs from MFNet dataset.

Chapter 3

Efficient object detection on embedded devices

Contents



Object detection

This chapter introduces our methods to build an efficient object detection system on embeded devices.

Precise and rapid object detection is the foundation of reliable video surveillance. However, the available computing resources under embedded environments is quite limited. In this chapter, we start by listing some best practices we experienced for training object detection models; then we provide implementation details of our two proposed approaches named **Mutual Guidance** and **PDF-Distil**, for improving the detection precision of existing models without introducing additional calculations and for reducing detection models' computational complexity, respectively; Finally we report our experimental results on two public object detection datasets to verify the effectiveness of the proposed practices and methods. Note that our two proposed methods are for general object detection, and are therefore not specific to multispectral scene analysis. The source code and pre-trained models of all mentioned methods in this chapter are publicly available at: https://github.com/ZHANGHeng19931123/MutualGuide.

This chapter concerns the following publications:

"Localize to Classify and Classify to Localize: Mutual Guidance in Object Detection", in *15th Asian Conference on Computer Vision (ACCV2020)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

"PDF-Distil: including Prediction Disagreements in Feature-based Distillation for object detection", in *32nd British Machine Vision Conference (BMVC2021)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

3.1 Best practices for training object detection models

Inspired by (Z. Zhang et al., 2019), we would like to explore best practices that apply to various object detection models. Specifically, we are interested in those training tricks and lightweight architectures that only decrease the training/inference speed by a small amount but can significantly improve the detection precision. For fair comparisons, all strategies are implemented within the same codebase. This codebase is initially based on a PyTorch implementation of SSD detector (W. Liu et al., 2016)[1], and is improved by gradually integrating newly proposed techniques into it. We found that maintaining this codebase up-to-date not only help us to evaluate the effectiveness of different SOTA methods, but also deepened our understanding of the entire object detection pipeline.

Algorithm 3.1 PyTorch-style pseudocode for training object detection with Mixup.

```
# generate mix ration via beta distribution
alpha = 1.0
lam = random.beta(alpha, alpha)
index = random.permutation(batch size)
# mix two images
images = lam*images + (1-lam)*images[index, :]
# mix two targets
targets_a, targets_b = targets, [targets[index[i]] for i in range(batch size)]
# model forward propagation
outputs = model(images)
# calculate loss via interpolation
loss_a, loss_b = criterion(outputs, targets_a), criterion(outputs, targets_b)
loss = lam * loss_a + (1 - lam) * loss_b
```

Mixup. Mixup (H. Zhang et al., 2018) is a data augmentation strategy originally proposed for image classification tasks. The main idea of Mixup is to linearly mix two images and their respective one-hot labels in a certain ratio. The value of this ratio is produced by the beta distribution. When adapting Mixup to the object detection task (Z. Zhang et al., 2019), we change the mix of one-hot labels into the mix of object detection losses. Algorithm 3.1 describes our implementation in a training iteration for object detection models.

[1] https://github.com/amdegroot/ssd.pytorch

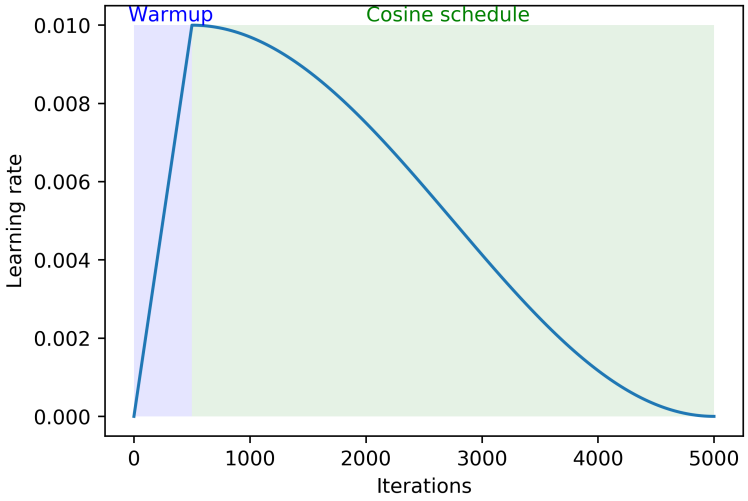


FIGURE 3.1 – An example of adjustment of learning rate when combining Warmup and Cosine annealing scheduler.

Warmup. This strategy of setting the learning rate hyperparameter for neural network training was firstly proposed by (He et al., 2016). Concretely, they apply a small learning rate (1e-2) for the first few hundred iterations to accelerate the training convergence, then the learning rate is resumed back to the base learning rate of 1e-1. Following this principle, we linearly increase the learning rate from a small value (1e-6 in our implementation) to the base value (1e-2) within the first 500 training iterations.

Cosine annealing schedule. Cosine annealing schedule is another strategy of setting the learning rate for neural network training. It consists in adjusting the learning rate according to the value of the cosine function, ranging from 0 to π. Compared to the normal step schedule, cosine learning rate decay is smoother. Specifically, the learning rate reduction is slow first, then fast, and finally slow down to 0. Figure 3.1 shows an example of the adjustment of learning rate when combining Warmup and Cosine annealing schedule.

Fusing BN layer into Convolutional layer. Batch Normalization (BN) layers (Ioffe & Szegedy, 2015) are widely adopted in today's deep neural networks. They help to accelerate the training loss convergence and leads to better accuracy. The common practice is to cascade a BN layer after a Convolutional layer. As shown in Figure 3.2, the convolutional layer and the BN layer can be merged to simplify the model architecture during inference time.

Context Enhancement Module. CEM was proposed in ThunderNet (Qin et al., 2019) as a light module for encoding global context information. It enlarges the receptive field by combining local features with global context. Its architecture is illustrated in Figure 3.3. As is shown, the introduced calculation is quite marginal (only 3 layers of 1x1 convolution).

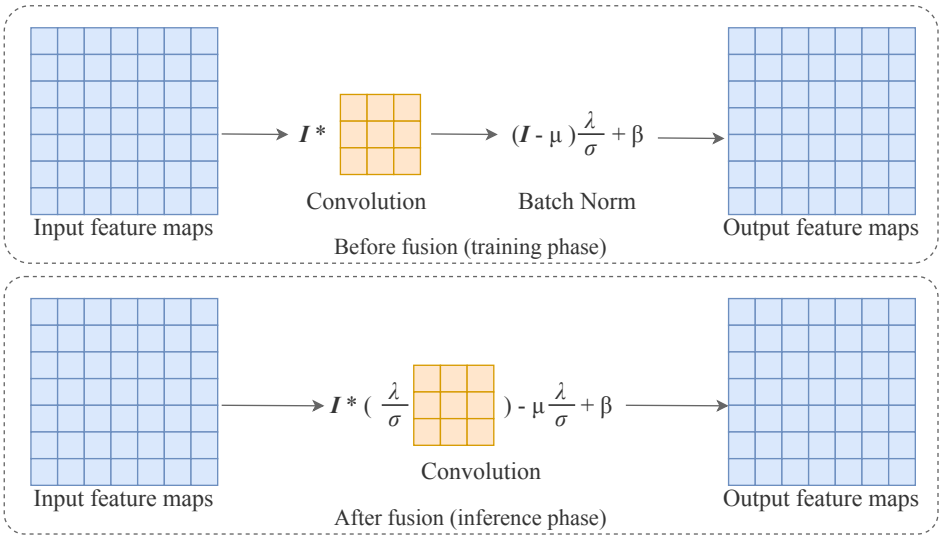


FIGURE 3.2 – Fusing batch norm layer into convolutional layer.

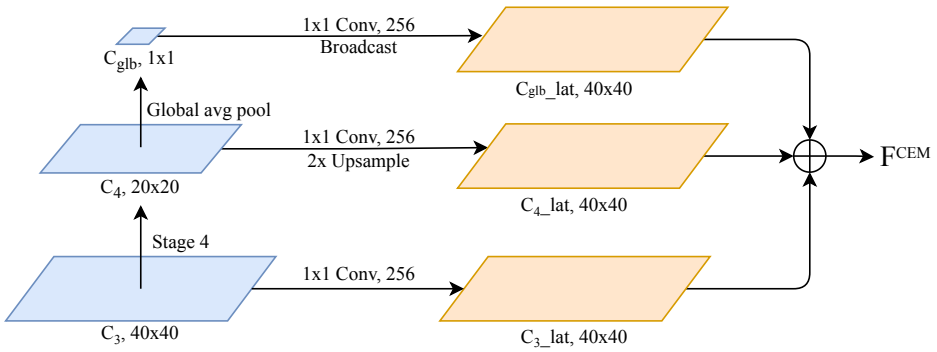


FIGURE 3.3 – Architecture of Context Enhancement Module (CEM).

Balanced L1 loss. The Smooth L1 loss (Girshick, 2015) is the usual function to optimize the bounding-box regression task in object detection. It is defined as:

$$Smooth\_L1(x) = \begin{cases} 0.5x^2 & if|x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{3.1}$$

The authors of Libra RCNN (Pang et al., 2019) found that the hard samples for bounding-box regression contribute to more than 70% of the gradients, which makes the model more sensitive to outliers. Derived from the conventional Smooth L1 loss, they proposed the Balanced L1 loss, in which the contribution of inliers is largely enhanced. Formally, they define Balanced L1 loss as:

$$Balanced\_L1(x) = \begin{cases} \frac{\alpha}{\beta}(\beta|x| + 1)\ln(\beta|x| + 1) - \alpha|x| & if|x| < 1 \\ \gamma|x| + C & otherwise \end{cases} \tag{3.2}$$

where α, β and γ are hyperparameters set as 0.5, 1.0 and 1.5 by default, respectively.

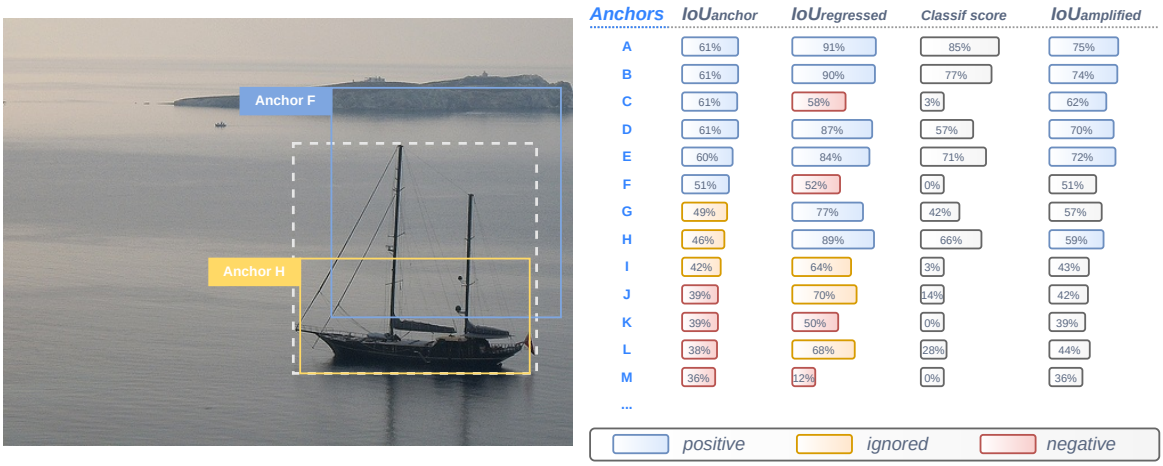


FIGURE 3.4 – Illustration of different anchor matching strategies for the boat image, resorting to IoU_{anchor} (static), $IoU_{regressed}$ (Localize to Classify) and $IoU_{amplified}$ (Classify to Localize). Anchors A-M are predefined anchor boxes around the boat in the picture (only F and H are visualized for the sake of clarity). Better viewed in colour.

3.2 Mutual Guidance for Anchor Matching

As mentioned in Section 2.1, in order to train an anchor-based object detection model, the predefined anchors should be assigned as positive ("it is a true object") or negative ("it is a part of the background") according to the matching between the anchor boxes and the ground truth boxes. Consequently, the bounding box regression loss is optimized with respect to the assigned positive anchors, and the instance classification loss is optimized with respect to the assigned positive as well as the negative anchors.

For the conventional static anchor matching strategy, the IoU between the predefined anchor boxes (i.e., regardless of any predictions) and the ground truth boxes (IoU_{anchor}) is the usual matching criterion. As shown in the IoU_{anchor} column of Figure 3.4, anchors with more than 50% of IoU_{anchor} are labelled as "positive", those with less than 40% of IoU_{anchor} are labelled as "negative", the rest are "ignored anchors". Note that at least one anchor should be assigned as positive, hence if there is no anchor with more than 50% of IoU_{anchor}, the anchor with the highest IoU_{anchor} is considered.

As previously explained in Section 2.1, such a static matching strategy is not content/context-sensitive and causes the task-misalignment problem. To tackle these two constraints, we propose a **Mutual Guidance** anchor matching mechanism. In particular, we constrain anchors that are well-localized to also be well-classified (**Localize to Classify**), and those well-classified to also be well-localized (**Classify to Localize**).

3.2.1 Localize to Classify

If an anchor is capable to precisely localize an object, this anchor must cover a good part of the semantically important area of this object and thus could be considered as an appropriate positive sample for classification. Drawing on this, we propose to leverage the IoU between the regressed bounding boxes (i.e., the network's localization predictions) and the ground truth boxes (noted as $IoU_{regressed}$) to better assign the anchor labels for classification. Inspired by the usual IoU_{anchor}, we compare $IoU_{regressed}$ to some given thresholds (discussed in the next paragraph) and then define anchors with $IoU_{regressed}$ greater than a high threshold as positive samples, and those with $IoU_{regressed}$ lower than a low threshold as negative samples (see $IoU_{regressed}$ column of Figure 3.4).

We now discuss a dynamic solution to set the aforementioned high and low thresholds. A fixed threshold (e.g., 50% or 40%) does not seem optimal since the object detection model's localization ability gradually improves during the training procedure and so does the $IoU_{regressed}$ for each anchor, leading to the assignment of more and more positive anchors which destabilizes the training. To address this issue, we propose a dynamic thresholding strategy: even though the IoU_{anchor} is not the best choice to accurately indicate the matching quality between anchors and objects, the number of assigned positive and ignored anchors does reflect the global matching conditions (brought by the size and the aspect ratio of the objects to detect), thus these numbers could be considered as reference values for our dynamic criterion. As illustrated in Figure 3.4, while applying the IoU_{anchor}-based anchor matching strategy with the thresholds being 50% and 40%, the number of positive anchors (N_p) and ignored anchors (N_i) are noted ($N_p = 6$ and $N_i = 3$ for the boat). We then use these numbers to label the N_p highest $IoU_{regressed}$ anchors as positive, and the following N_i anchors as ignored. More formally, we exploit the N_p-th largest $IoU_{regressed}$ as the high threshold, and the $(N_p + N_i)$-th largest $IoU_{regressed}$ as the low threshold. Using this, our **Localize to Classify** matching strategy evolves with the network's localization capacity and maintains a consistent number of anchor samples assigned to both categories (positive/negative) during the whole training procedure.

3.2.2 Classify to localize

The positive anchor samples in **Classify to Localize** are assigned according to the model's classification predictions (noted as p). Specifically, p is the predicted classification score for the corresponding object category, e.g., the $Classifscore$ column of Figure 3.4 indicates the classification score p for the boat category. Nevertheless, this p is not effective enough to be used directly for assigning good positive anchors for the bounding box regression optimization. It is especially true at the beginning of the training process, when the network's weights are almost random values and all predicted classification scores are close to zero. The $IoU_{regressed}$ is optimized on the basis of the IoU_{anchor}, therefore we have $IoU_{regressed} \geq IoU_{anchor}$ in most cases (even at the beginning of the training), and this property helps to avoid such a cold

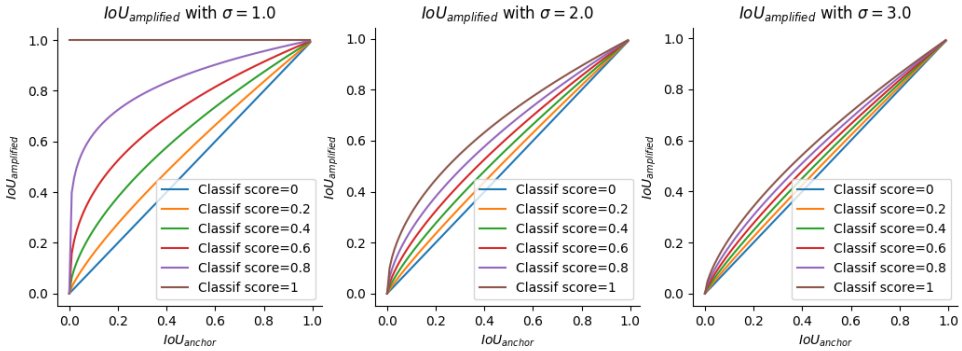


FIGURE 3.5 – Illustration of $IoU_{amplified}$ with different σ values (from left to right signifies 1, 2 or 3). $IoU_{amplified}$ equals to IoU_{anchor} when $p = 0$.

start problem and ensures training stability. Symmetrically to the **Localize to Classify** strategy, we now propose a **Classify to Localize** strategy based on $IoU_{amplified}$ defined as:

$$IoU_{amplified} = (IoU_{anchor})^{\frac{\sigma - p}{\sigma}} \tag{3.3}$$

where σ is a hyperparameter aiming at adjusting the degree of amplification. Inspired by the Focal Loss (Lin, Goyal, et al., 2017), we chose Equation 3.3 as the simplest one able to amplify the IoU of anchors according to the correct classification predictions p. Its behaviour is shown in Figure 3.5. The $IoU_{amplified}$ is always higher than the IoU_{anchor}, and the amplification is proportional to the predicted p. In particular, the amplification is stronger for smaller σ (note that σ should be larger than 1), and disappears when σ becomes large.

Similarly to the **Localize to Classify** strategy, we apply a dynamic thresholding strategy to keep the number of assigned positive samples for the localization task and for the classification task consistent, e.g., we assign in Figure 3.4, the top N_p anchors with the highest $IoU_{amplified}$ as positive samples. Note that there is no need for selecting ignored or negative anchors for the localization task, since the background does not have an associated ground truth box.

As discussed in Section 2.1, IoU_{anchor} is not sensitive to the content or the context of an object. Our proposed **Localize to Classify** and **Classify to Localize**, however, attempt to adaptively label the anchor samples according to their visual content and context information. Considering anchor F and anchor H in Figure 3.4, one can tell that anchor H is better than anchor F for recognizing this boat, even with a smaller IoU_{anchor}. Using both our strategies, anchor H has been promoted as positive thanks to its excellent prediction quality on both tasks, whereas anchor F has been labelled as negative even though it has a large IoU_{anchor}.

3.2.3 Discussion on the task-misalignment problem

Since **Localize to Classify** and **Classify to Localize** are independent strategies, they could possibly assign contradictory positive/negative labels (e.g, the anchor C in Figure 3.4 is labelled negative for the classification task but positive for the bounding box regression task). This happens when one anchor entails a good prediction on one task and a poor prediction on the other (i.e. they are misaligned predictions). Dealing with such contradictory labels, as we do with **Mutual Guidance**, does not harm the training process. On the contrary, our method tackles the task-misalignment problem since the labels for one task are assigned according to the prediction quality on the other task, and vice versa. This mechanism forces the network to generate aligned predictions: If the classification prediction from one anchor is good while its localization prediction is bad, the **Mutual Guidance** will give a positive label on the localization task to this anchor, to constrain it to be better at localizing as well while giving a negative label (i.e. background) on the classification task to avoid misaligned predictions. In fact, the predicted classification score of this mislocalized anchor should be low enough for the anchor to be suppressed by the Non-Maximum Suppression (NMS) procedure[2] during inference. The same reasoning holds for a good localization prediction with a bad classification one.

On the contrary, if a network always assigns similar positive/negative labels (as done in standard static methods) to both tasks during training, one cannot guarantee that there will be no misalignment of the localization and the classification predictions at inference time. Keeping anchors (after NMS) with misaligned predictions is especially harmful for strict evaluation metrics such as AP75.

3.3 Prediction Disagreement aware Feature Distillation

As explained in Section 2.3, knowledge distillation is an effective method to reduce the complexity of object detection models. Figure 2.5 (A) illustrates that the foreground-background imbalance problem is the major obstacle for efficient knowledge transfer in the detection distillation problem. We tackle this imbalance problem from a sampling perspective, and we propose to include the teacher-student prediction disagreements into the knowledge distillation (KD) framework.

3.3.1 Prediction disagreement aware feedback branch

Specifically, we illustrate the overview of the proposed **PDF-distil** method in Figure 3.6. The student model employs a simpler network architecture than the teacher model, namely thinner or shallower backbone and neck networks in the context of object detection. Note that in Figure 3.6 the multiscale detection architecture (Lin, Dollár, et al., 2017; S. Liu et al., 2018) is not presented for the sake of clarity. The yellow blocks in Figure 3.6 show that the training of the student model is supervised

[2]Non-Maximum Suppression (NMS) is an operation to remove redundant bounding boxes from predictions, more information can be found in https://www.pyimagesearch.com/2014/11/17/non-maximum-suppression-object-detection-python/.

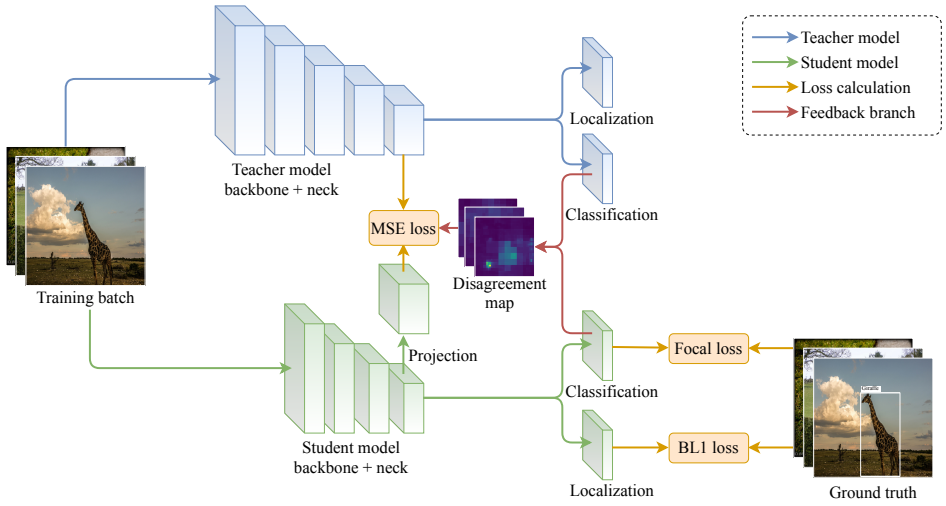


FIGURE 3.6 – Overview of the proposed PDF-distil method. We have added a prediction disagreement aware feedback branch in a traditional feature-based detection distillation framework.

by the normal object detection loss (including the instance classification and the bounding-box regression losses) as well as the knowledge transfer loss, which is defined as the Mean Square Error (MSE) between the intermediate feature maps of the teacher model and the projected feature maps of the student model. Following the common practice, this projection is performed through a 1×1 convolution to map the student hidden layer to the teacher hidden layer.

The main contribution of **PDF-distil** consists in adding a prediction disagreement aware feedback branch in a traditional feature-based detection distillation framework. This feedback branch leverages the prediction difference between the teacher and student to generate a disagreement map, which is used as a weighting mask applied to the knowledge transfer loss. Our intuition is that regions where the two models make different object detection predictions are actually regions where the student model struggles the most. Thus, enhancing distillation loss on these regions could greatly reduce the performance gap between both models.

3.3.2 Disagreement mapping

In order to obtain the aforementioned disagreement map, we compute the distance between the respective classification branches of the teacher model and the student model[3]. Formally, let C^t and C^s respectively represent the output probability distributions from the classification branches of the teacher model and the student model, let N denotes the number of object categories. Assume that there are M classification predictions associated to a specific feature map location. To be more specific, for anchor-based methods, M equals to the number of anchors per location, e.g., $M = 6$ for SSD (W. Liu et al., 2016) and $M = 9$ for RetinaNet (Lin, Goyal, et al., 2017); for anchor-free methods like FCOS (Tian et al., 2019), M equals to 1 since each

[3]Since localization predictions on background areas are meaningless, we do not consider this prediction difference in the localization branches.

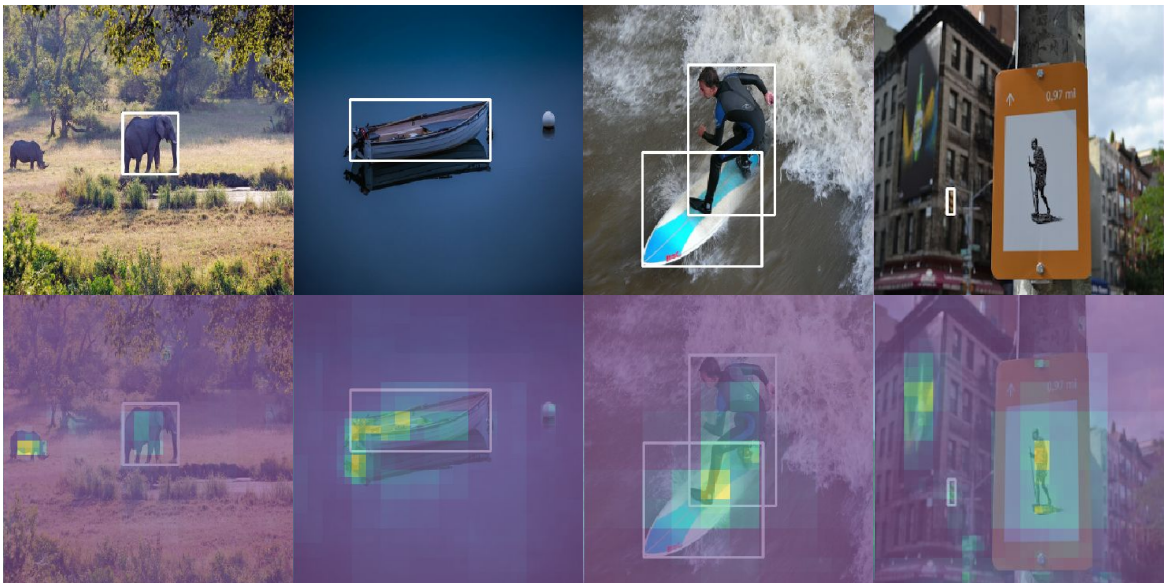


FIGURE 3.7 – Visualization of our PDF-Distil sampling strategies for feature-based detection distillation. Note that for better comparisons with previous methods, we show the same examples as in Figure 2.5.

feature map location only produces one bounding-box prediction. The prediction disagreement at each feature location ($D_{h,w}$) (w and h are the coordinates of feature location) is defined as:

$$D_{h,w} = \sum_{M}\sum_{N} \mathcal{F}(C^{t}_{h,w}, C^{s}_{h,w}) \tag{3.4}$$

where $\mathcal{F}$ is a given dissimilarity function (in Section 3.4, we compare KL-divergence, L1 and L2 distances). Let H, W and C denote the height, width and depth of the feature maps, the actual weighting value at each location on the disagreement map ($M_{h,w}$) is assigned as:

$$M_{h,w} = \frac{H \times W \times D_{h,w}}{\sum_{H}\sum_{W} D_{h,w}} \,. \tag{3.5}$$

As shown in Figure 3.7, the generated disagreement map is biased towards "hard" regions, such as unknown objects (first example), reflection in water (second example), object junctions (third example) and ambiguous objects (fourth example).

Let X^t denote the output feature maps of the teacher backbone network and X^s the projected feature maps from the student model, the weighted knowledge transfer loss L_{kd} (here kd represents knowledge distillation) is computed as:

$$L_{kd} = \frac{\sum_{H}\sum_{W}(M_{h,w} \times \sum_{C}(X^t - X^s)^2)}{H \times W \times C} \,. \tag{3.6}$$

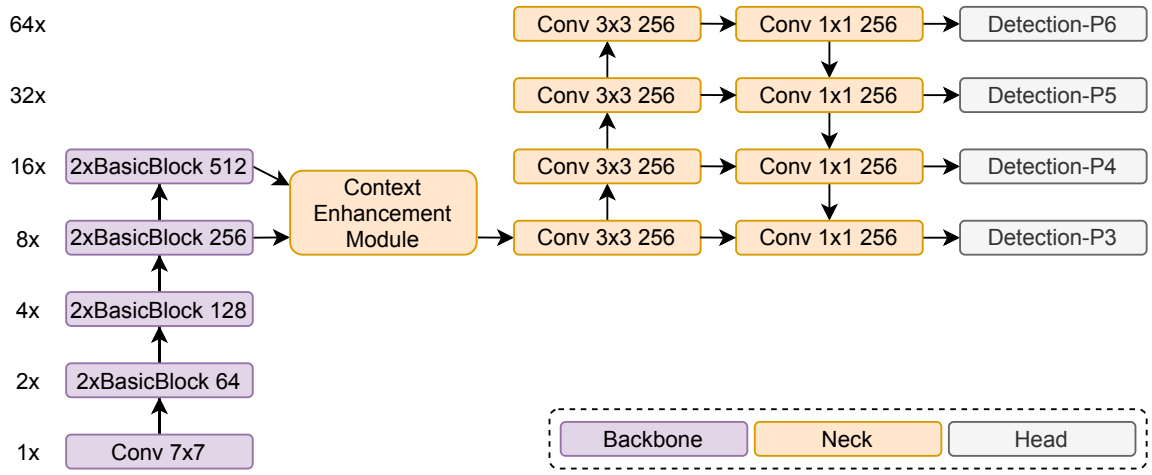


FIGURE 3.8 – Architecture of the implemented RetinaNet.

3.4 Experimental results

3.4.1 Implementation details

Network architectures. In order to test the effectiveness of our proposed **Mutual Guidance** (H. Zhang, Fromont, Lefèvre, et al., 2020) and **PDF-distil** methods, we conduct experiments on the single-stage object detectors RetinaNet (Lin, Goyal, et al., 2017) with the widely-used ResNet-18 (He et al., 2016) backbone network. Its detailed network architecture is illustrated in Figure 3.8. As is shown, multiscale feature maps are extracted via ResNet-18 backbone, then sent to FPN (Lin, Dollár, et al., 2017) neck for generating high-semantic feature pyramid. A specific RetinaNet detection head is attached to each pyramid level, for classifying and localizing objects within a certain scale. Specifically, upper-level detection heads are responsible for detecting large objects, while lower-level detection heads handle small object detection.

In order to realize knowledge distillation, we need to construct a more precise object detector as the teacher model. Thus, we replace the "shallower" ResNet-18 (He et al., 2016) backbone by the deeper ResNet-34, and replace the simpler FPN neck by the more complex PAFPN (S. Liu et al., 2018). To compare the computational cost of the teacher and student models, we summarize in Table 3.1 the amount of learnable parameters (denoted as *Params*) and the amount of Multiply–Accumulate operations (denoted as *MACs*) for both models[4]. It can be observed that the student requires much less computing resources than the teacher.

Parameter initialization. For all object detection models, the backbone network (ResNet-18 or ResNet-34) is pre-trained on the ImageNet-1k dataset (Deng et al., 2009), while the other parts of the networks (i.e., the neck and the head networks) are randomly initialized according to the method described in (He et al., 2015).

[4]These values are measured via https://github.com/Lyken17/pytorch-OpCounter.

Model	*Params*	*MACs*
Teacher	4.01e+07	3.86e+10
Student	3.00e+07	2.35e+10

TABLE 3.1 – Comparison between the computational cost of the teacher model and student model.

Model	Best practices	*mAP*	*AP*50	*AP*75
RetinaNet with ResNet-18 backbone	Baseline	49.7	76.6	53.3
	+CEM	50.4 (+0.7)	77.8	53.7
	+Mixup&Warmup&Cosine	52.9 (+3.2)	80.1	57.0
	+Balance L1 loss	**54.6 (+4.9)**	**79.4**	**58.6**

TABLE 3.2 – Precision improvements when integrating different best practices. Experiments are conducted on the PASCAL VOC dataset. The best score combining all best practices is in bold.

Data preprocessing. We adopt multiscale training and single-scale evaluation, i.e., the input image resolution is randomly resized to 256×256 or 320×320 or 384×384 during training phase and fixed to 320×320 during evaluation phase. Several data augmentation strategies are applied during the training phase, such as random image flipping, shifting, cropping, padding, noising and Mixup (H. Zhang et al., 2018) (as explained in Section 3.1).

Network optimization. We use the Stochastic Gradient Descent (SGD) optimizer with 16 images per mini-batch and with an initial learning rate of 1e-2. The Warmup strategy (Goyal et al., 2017) is applied to stabilize the training at the beginning, followed by the cosine annealing strategy (Loshchilov & Hutter, 2016) for learning rate decay. Models are trained for 70 and 140 epochs for PASCAL VOC and MS COCO, respectively. We use Balanced L1 loss (Pang et al., 2019) and Focal Loss (Lin, Goyal, et al., 2017) to optimize the localization and the classification branch of RetinaNet.

3.4.2 Results for best practices

This section elaborates on the influence of the different practices presented in Section 3.1. As illustrated in Table 3.2, the detection precision is gradually improved by integrating the presented practices. Specifically, the implementation of Balance L1 loss, Warmup&Cosine annealing, and Mixup are only involved in the training phase of object detection models. Context Enhancement Module (CEM) yields the modification on the model architecture, but the introduced additional parameters and calculation is trivial.

3.4.3 Results for MutualGuidance

Note that all the following object detection experiments implement the aforementioned **best practices** to optimize object detection models.

Model	Matching strategy	mAP	$AP50$	$AP75$
RetinaNet with ResNet-18 backbone	IoU_{anchor}-based	54.6	79.4	58.6
	Localize to Classify	56.1 (+1.5)	80.1	61.0
	Classify to Localize	55.1 (+1.0)	79.8	59.2
	Mutual Guidance	**56.5 (+1.9)**	**80.1**	**61.5**

TABLE 3.3 – Comparison of different anchor matching strategies (the usual IoU_{anchor}-based, proposed Localize to Classify, Classify to Localize and Mutual Guidance) for object detection. Experiments are conducted on the PASCAL VOC dataset. The best score is in bold.

Model	Matching strategy	mAP	$AP50$	$AP75$	AP_s	AP_m	AP_l
RetinaNet with ResNet-18 backbone	IoU_{anchor}-based	33.7	51.5	35.5	15.3	39.5	48.4
	Mutual Guidance	**35.0 (+1.3)**	**52.2**	**37.2**	**16.4**	**40.5**	**50.5**

TABLE 3.4 – AP performance for object detection on MS COCO dataset using 2 different anchor matching strategies: the usual IoU_{anchor}-based one and our complete approach marked as Mutual Guidance.

3.4.3.1 Precision improvements

PASCAL VOC dataset. We evaluate the effectiveness of both components (**Localize to Classify** and **Classify to Localize**) of **Mutual Guidance** w.r.t. the usual IoU_{anchor}-based matching strategy when applied on the same deep learning architectures. The results obtained on the PASCAL VOC dataset are given in Table 3.3. Both proposed anchor matching strategies consistently boost the performance of the "vanilla" networks, and their combination (**Mutual Guidance**) leads to the best AP and all other evaluation metrics.

In particular, we observe that the improvements are small on AP50 (around 0.5%) but significant on AP75 (around 3%), which means that we obtain more precise detections. As analysed in Section 2.1, this comes from the task-misalignment problem faced with the usual static anchor matching methods. This issue leads to retain well-classified but poorly-localized predictions and suppress well-localized but poorly-classified predictions, which in turns results in a significant drop of the AP score at strict IoU thresholds, e.g., AP75. In **Mutual Guidance**, however, training labels for one task are dynamically assigned according to the prediction quality on the other task and vice versa. This connection makes the classification and localization tasks consistent along the training phase, as such avoids this task-misalignment problem.

We also notice that **Localize to Classify** alone brings a higher improvement than **Classify to Localize** alone. We hypothesize two possible reasons for this: 1) most object detection errors come from wrong classifications instead of imprecise localizations, so the classification task is more difficult than the localization task and thus, there is more room for the improvement on this task; 2) the amplification proposed in Equation 3.3 may not be the most appropriate one to take advantage of the classification task for optimizing the bounding box regression task.

FIGURE 3.9 – Visualization of the difference in the label assignment during training phase. White dotted-line boxes represent ground truth boxes; Red anchor boxes are assigned as positive by IoU_{anchor}-based strategy, while considered as negative or ignored by Localize to Classify or Classify to Localize; Green anchor boxes are assigned as positive by Localize to Classify but negative or ignored by IoU_{anchor}-based; Yellow anchor boxes are assigned as positive by Classify to Localize but negative or ignored by IoU_{anchor}-based strategy.

MS COCO dataset. We then conduct experiments on the more difficult MS COCO dataset and report our results in Table 3.4. Note that according to the scale range defined by MS COCO, APs of small, medium and large objects are listed. In this dataset also, our **Mutual Guidance** strategy consistently brings some performance gains compared to the IoU_{anchor}-based baselines. We notice that our AP gains on large objects is significant (around 2%). This is because larger objects generally have more matched positive anchors, which offers more room for improvements to our method. Since the **Mutual guidance** strategy only involves the training phase, and since there is no difference between IoU_{anchor}-based and our method during the evaluation phase, these improvements can be considered cost-free.

3.4.3.2 Qualitative analysis

Label assignment visualization. Here, we would like to explore the reasons for the performance improvements by visualizing the difference in the label assignment between the IoU_{anchor}-based strategy and the **Mutual Guidance** strategy during training. According to the examples shown in Figure 3.9, we can conclude that the IoU_{anchor}-based strategy only assigns the "positive" label to anchors with sufficient IoU with the ground truth box, regardless of their content or context, whereas our proposed **Localize to Classify** and **Classify to Localize** strategies dynamically assign "positive" labels to anchors covering semantic discriminant parts of the object (e.g., upper body of a person, main body of animals), and assign "negative" labels to anchors with complex background, occluded parts, or anchors containing nearby objects. We believe that our proposed instance-adaptive strategies make the label assignment more reasonable, which is the main reason for performance increase.

Detection results visualization. Figure 3.10 illustrates on a few images from the PASCAL VOC dataset the different behaviours shown by our **Mutual Guidance** method and the baseline anchor matching strategy. As analysed in Section 2.1, we can find misaligned predictions (good at classification but poor at localization) from IoU_{anchor}-based anchor matching strategy. As shown in the figure, our method gives better results when different objects are close to each other in the image, e.g. "man riding a horse" or "man riding a bike". With the usual IoU_{anchor}-based anchor matching strategy, the instance localization and classification tasks are optimized independently of each other. Hence, it is possible that, during the evaluation phase, the classification prediction relies on one object whereas the bounding box regression targets the other object. However, such a problem is rarer with the **Mutual Guidance** strategy. Apparently, our anchor matching strategies introduce interactions between both tasks and make the predictions of localization and classification aligned, which substantially eliminated such false positive predictions.

3.4.4 Results for PDF-Distil

Note that all the following knowledge distillation experiments implement both **best practices** and **Mutual Guidance** to optimize object detection models.

3.4.4.1 Ablation study

Ablation experiments are conducted on PASCAL VOC to explore the relationship between the teacher-student prediction disagreements and the knowledge transfer effects. In Table 3.5, we consider eight different feature sampling strategies for detection distillation: 1) the baseline setting where all samples are treated equally (equivalent to Fitnets (Romero et al., 2015)); 2-5) hard sampling strategies where the distillation is only conducted on 25% or 50% of feature areas with the most similar or the most different teacher-student predictions; 6-8) the proposed adaptive sampling approach with respectively KL-divergence, L1 distance or L2 distance as the dissimilarity function in Equation 3.4. The results are summarized in Table 3.5. When comparing the distillation results of the four hard sampling strategies (i.e., 2-5), we can conclude that feature samples with different teacher-student predictions

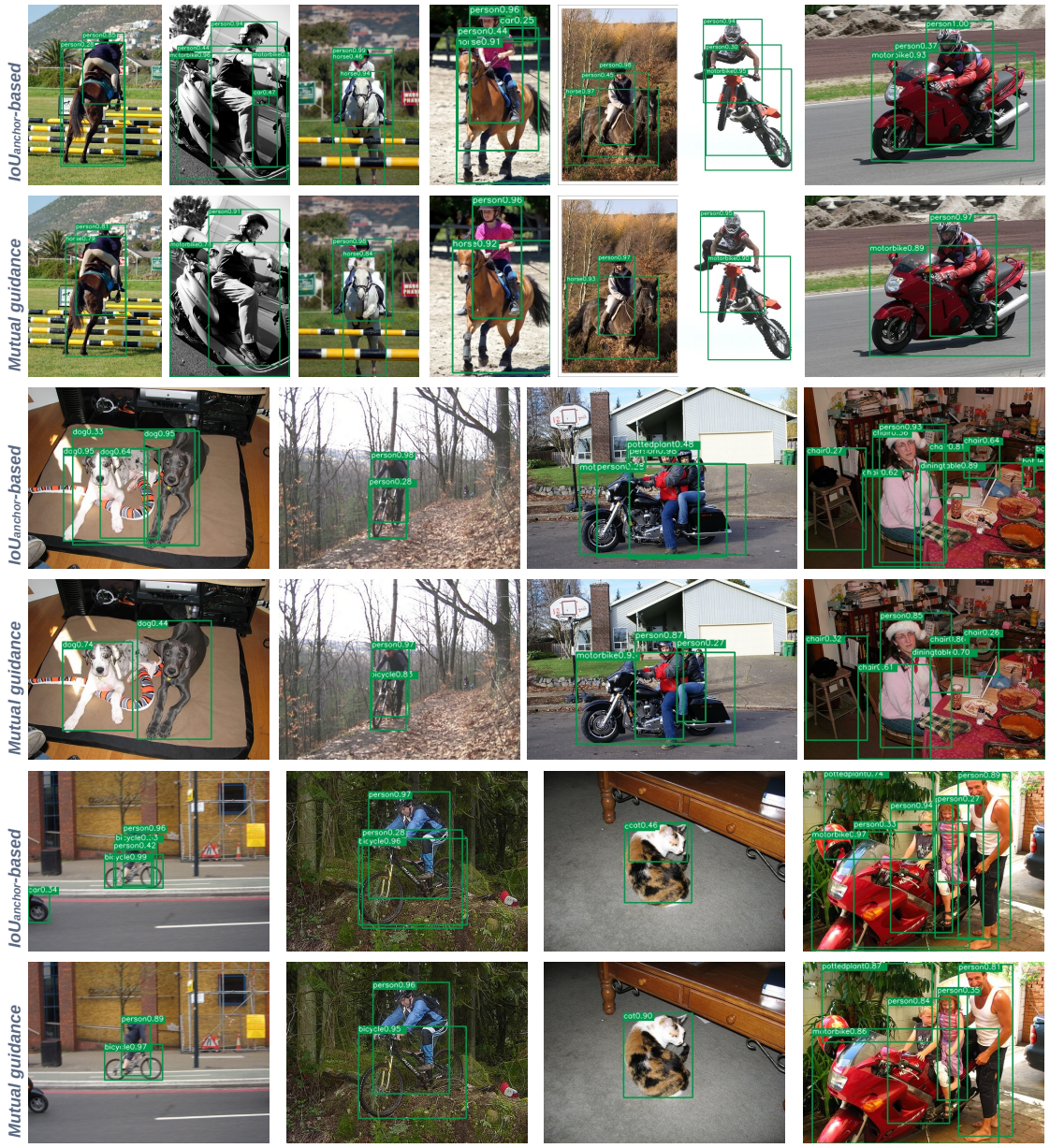


FIGURE 3.10 – Examples of detection results using the IoU_{anchor}-based anchor matching strategy (odd lines) and our proposed **Mutual Guidance** one (even lines). The results are given for all images after applying a Non-Maximum Suppression process with a IoU threshold of 50%.

are much more effective than those with similar predictions. This finding validates our initial hypothesis that the disagreements between the teacher-student object detection predictions can be regarded as an indicator of the importance for feature-based distillation. Moreover, regardless of the specific dissimilarity function, the adaptive sampling strategies (6-8) outperform the hard sampling strategies (2-5), indicating the effectiveness of the proposed dynamic weighting mechanism. As for the selection of the dissimilarity function, L2 distance (i.e., 8) demonstrates a constant advantage for all backbone-neck-head combinations. Therefore, we choose L2 distance as the dissimilarity function for the following experiments.

	Models	*mAP*	*AP*50	*AP*75
Teacher	ResNet34-PAFPN-RetinaNet-MG	60.0	82.5	65.3
Student	ResNet18-FPN-RetinaNet-MG	56.5	80.1	61.5
	1) All samples equally	57.8 (+1.3)	81.4	62.9
	2) 25% most similar predictions	57.5 (+1.0)	81.0	62.6
	3) 50% most similar predictions	57.8 (+1.3)	81.3	62.9
	4) 50% most different predictions	59.2 (+2.7)	82.3	64.6
	5) 25% most different predictions	59.3 (+2.8)	82.3	64.8
	6) PDF-Distil (KL-divergence)	59.4 (+2.9)	82.5	64.7
	7) PDF-Distil (L1 distance)	59.5 (+3.0)	82.7	65.3
	8) PDF-Distil (L2 distance)	**59.8 (+3.3)**	**83.0**	**65.4**

TABLE 3.5 – Ablation studies on PASCAL VOC. We compare eight different feature sampling strategies for detection distillation, and the proposed PDF-Distil with L2 distance as the dissimilarity function achieves the best result.

	Models	*mAP*	*AP*50	*AP*75	D_{pred}
Teacher	ResNet34-PAFPN-RetinaNet-MG	60.0	82.5	65.3	-
Student	ResNet18-FPN-RetinaNet-MG	56.5	80.1	61.5	2.96E-4
	w/ Fitnets (Romero et al., 2015)	57.8 (+1.3)	81.4	62.9	2.54E-4
	w/ Fine-grained (T. Wang et al., 2019)	58.6 (+2.1)	81.6	64.4	2.66E-4
	w/ Decoupled (Guo et al., 2021)	58.4 (+1.9)	81.8	63.5	2.43E-4
	w/ Prime-aware (Y. Zhang et al., 2020)	58.6 (+2.1)	81.9	63.7	2.44E-4
	w/ PDF-Distil (L2 distance)	**59.8 (+3.3)**	**83.0**	**65.4**	**2.20E-4**

TABLE 3.6 – Comparisons with SOTA detection distillation methods on PASCAL VOC.

	Models	*mAP*	*AP*50	*AP*75	D_{pred}
Teacher	ResNet34-PAFPN-RetinaNet-MG	38.7	56.2	41.4	-
Student	ResNet18-FPN-RetinaNet-MG	35.0	52.2	37.2	1.75E-4
	w/ Fitnets (Romero et al., 2015)	35.6 (+0.6)	52.7	37.8	1.62E-4
	w/ Fine-grained (T. Wang et al., 2019)	36.0 (+1.0)	53.0	38.3	1.58E-4
	w/ Decoupled (Guo et al., 2021)	35.9 (+0.9)	53.2	37.7	1.53E-4
	w/ Prime-aware (Y. Zhang et al., 2020)	35.6 (+0.6)	52.8	37.7	1.57E-4
	w/ PDF-Distil (L2 distance)	**36.9 (+1.9)**	**54.2**	**39.1**	**1.46E-4**

TABLE 3.7 – Comparisons with SOTA detection distillation methods on MS COCO.

3.4.4.2 Comparison with state-of-the-art

As shown in Tables 3.6 and 3.7, we further compare our method with SOTA detection distillation methods on PASCAL VOC and MS COCO datasets. The results show that for either backbone-neck-head combinations and on both datasets, our method outperforms all existing KD methods. In particular, our method brings more than 3% (respectively 2%) of absolute precision improvements in comparison

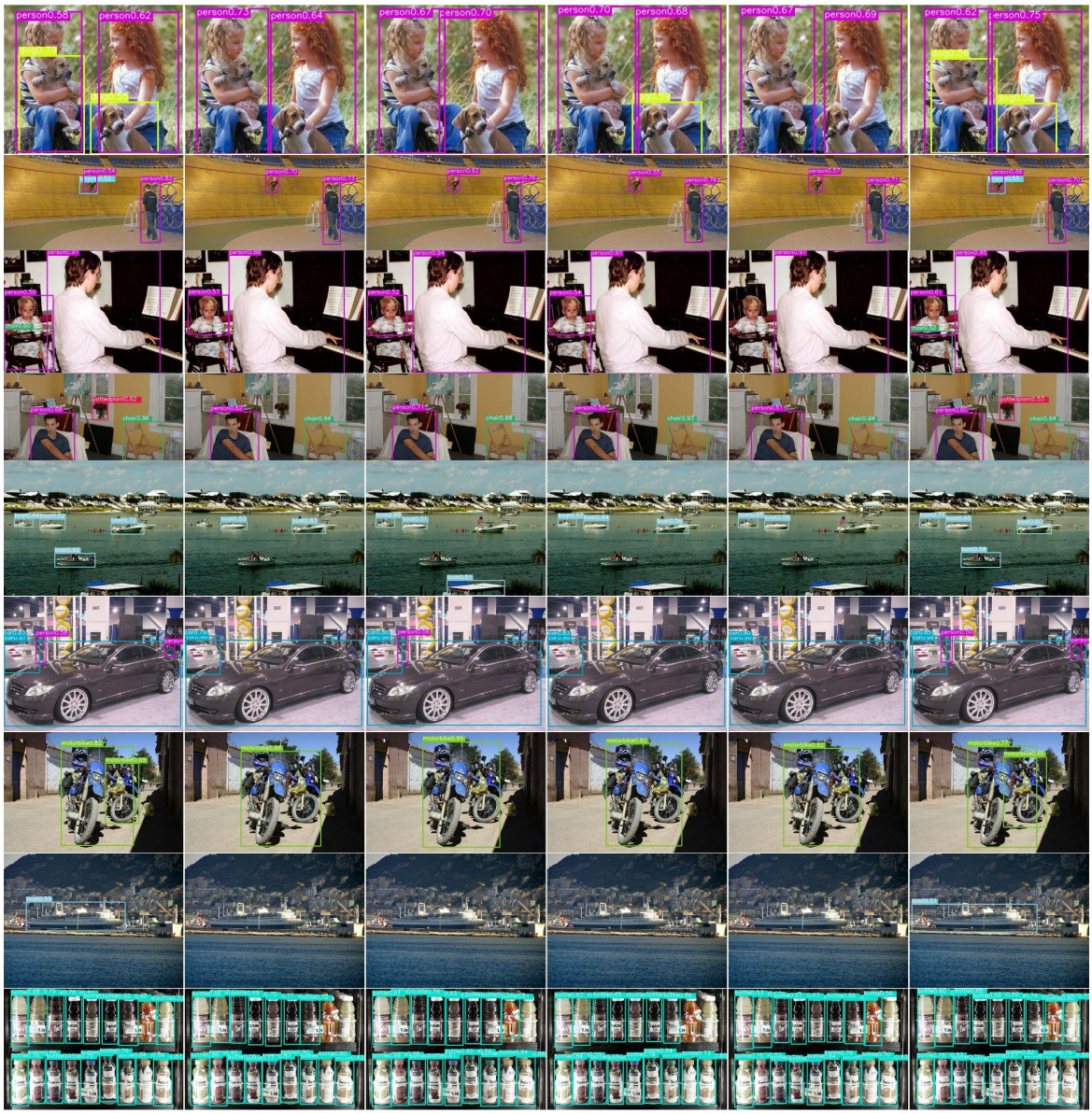


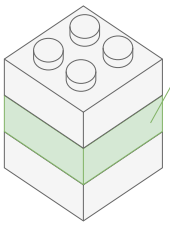
FIGURE 3.11 – Visualization of some detection results from teacher model and student models distilled by Fitnets, Fine-grained, Decoupled, Prime-aware and our PDF-Distil. Our method gives detection results more similar to the teacher model than the other methods.

to student models without KD on PASCAL VOC (resp. MS COCO), and about 1% of absolute improvements to all previous detection distillation methods. Moreover, we report on the test set of each dataset the absolute difference between the detection predictions of the teacher model and the student model (denoted as D_{pred}), and we notice that our method effectively reduces the teacher-student prediction difference. Figure 3.11 illustrates the detection results on a few images treated by the teacher model, Fitnets (Romero et al., 2015), Fine-grained (T. Wang et al., 2019), Decoupled (Guo et al., 2021), Prime-aware (Y. Zhang et al., 2020) and our method. As is shown, our method gives detection results more similar to the teacher model than the other SOTA methods that miss some objects (e.g., dog, bicycle, potted plant, chair).

Chapter 4

Information fusion from multispectral data

Contents



Multispectral fusion

This chapter introduces our methods for better information fusion from multispectral data.

Visible and thermal images are expected to be complementary when used for object detection applications. Therefore, combining both spectra can achieve more robust and accurate detection performance. When dealing with multispectral fusion, the major difficulty lies in cases where different cameras provide contradictory information, e.g., contradictory features or predictions. The three proposed methods aim to tackle this problem through different fusion strategies. Extensive experiments on public datasets demonstrate their effectiveness.

This chapter concerns the following publications:

"Multispectral Fusion for object detection with Cyclic Fuse-and-Refine blocks" in *27th International Conference on Image Processing (ICIP2020)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

"Guided Attentive Feature Fusion for Multispectral Pedestrian Detection", in *Winter Conference on Applications of Computer Vision (WACV2021)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

4.1 Multispectral Fusion with Cyclic Fuse-and-Refine

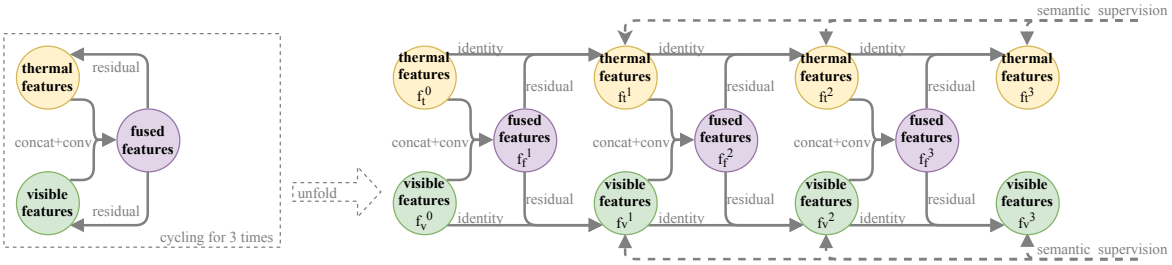


FIGURE 4.1 – Illustration (folded on the left part and unfolded on the right) of the proposed CFR fusion method with 3 loops.

In our first proposed method **Cyclic Fuse-and-Refine** (CFR), we seek to decrease the multispectral feature inconsistency via recursively refining the inconsistent (i.e., monospectral) features with the consistent (i.e., fused) features. The fusion operation aims at extracting concordant features from different spectra, and the refinement operation aims at reducing multispectral inconsistency. An illustration of the proposed module with 3 Fuse-and-Refine loops is presented in Figure 4.1. Such a fusion scheme has two advantages: 1) since the fused features are generally more discriminative than the monospectral ones, the refined spectral features should also be more discriminative than the original spectral features and the fuse-and-refine loop gradually improves the overall feature quality; 2) since the monospectral features keep being refined with the same features, their inconsistency is progressively reduced.

4.1.1 Fuse-and-Refine

To make our implementation as simple as possible, we use a concatenation operation for the fusion and an addition operation for the refinements. In each loop i, for the fused, thermal and visible features (subscript f, t and v, respectively), the multispectral feature fusion can be formalized as:

$$f_f^i = \mathcal{F}(\sigma(f_t^{i-1}, f_v^{i-1})) \tag{4.1}$$

As for the previous equations, σ is a feature concatenation operation, and $\mathcal{F}$ is a 3×3 convolution followed by a Batch Normalization (BN) layer. The fused features are then assigned as residuals of the spectral features for refinement:

$$\begin{aligned} f_t^i &= \mathcal{H}(f_t^{i-1} + f_f^i) \\ f_v^i &= \mathcal{H}(f_v^{i-1} + f_f^i) \end{aligned} \tag{4.2}$$

where $\mathcal{H}$ simply denotes the Rectified Linear Units (ReLU) function.

4.1.2 Cyclic computation

As mentioned earlier, the cyclic structure allows a progressive improvement of the overall multispectral feature quality and an iterative reduction of the multispectral feature inconsistency. Here are, for example, the computations done by the cyclic module with 3 loops:

$$\begin{aligned} f_f^1 &= \mathcal{F}(\sigma(f_t^0, f_v^0)), f_t^1 = \mathcal{H}(f_t^0 + f_f^1), f_v^1 = \mathcal{H}(f_v^0 + f_f^1) \\ f_f^2 &= \mathcal{F}(\sigma(f_t^1, f_v^1)), f_t^2 = \mathcal{H}(f_t^1 + f_f^2), f_v^2 = \mathcal{H}(f_v^1 + f_f^2) \\ f_f^3 &= \mathcal{F}(\sigma(f_t^2, f_v^2)), f_t^3 = \mathcal{H}(f_t^2 + f_f^3), f_v^3 = \mathcal{H}(f_v^2 + f_f^3) \end{aligned} \tag{4.3}$$

The features f_t^0 and f_v^0 are the original thermal and visible features, before any fusion process. We assume (and will show empirically) that the inconsistency between f_t and f_v gradually decreases after each Fuse-and-Refine loop.

4.1.3 Semantic supervision

In order to better guide the multispectral feature fusion, an auxiliary semantic segmentation task is used to bring separate supervision information for each refined monospectral features: after being refined with the fused features, the thermal and visible features go through a 1×1 convolution to predict pedestrian masks. We use the usual DICE loss (Dice, 1945) to supervise the prediction of the pedestrian masks, which is defined as:

$$L_{dice} = 1 - \frac{2\,|A \cap B|}{|A| + |B|} \tag{4.4}$$

where A represents the predicted pedestrian mask and B represents the ground truth pedestrian mask.

4.1.4 Final fusion

We aggregate all the refined monospectral features to generate the final fused features. The aggregation is a simple element-wise average function. Let I be the total number of loops, the final computation is:

$$f_{final} = \frac{1}{2I}\left(\sum_{i=1}^{I} f_t^i + \sum_{i=1}^{I} f_v^i\right) \tag{4.5}$$

Note that here we exclude the original features f_t^0 and f_v^0 because they generally have a fewer discriminative quality than the refined features and including them does not bring any precision improvement.

4.2 Progressive Spectral Fusion

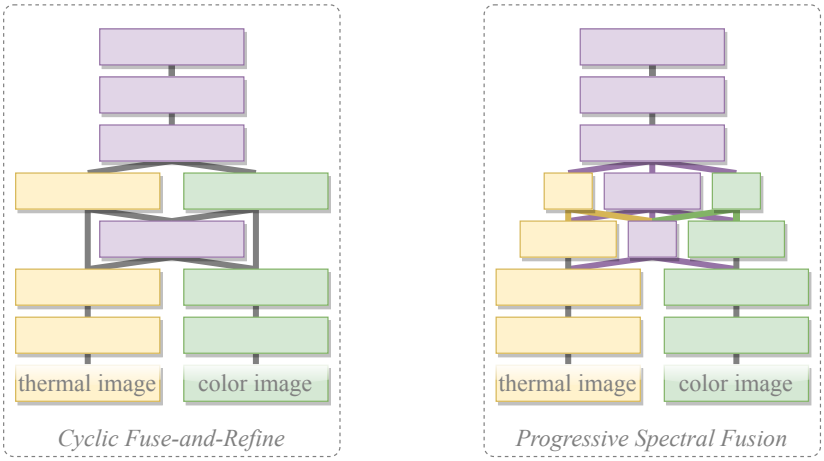


FIGURE 4.2 – Comparison between Cyclic Fuse-and-Refine (CFR) and Progressive Spectral Fusion (PS-Fuse).

Progressive Spectral Fusion (PS-Fuse) is our follow-up work on Cyclic Fuse-and-Refine (CFR), where we avoid the heavier scheme of recursive operations of CFR. As shown in Figure 4.2 right, we reduce the multispectral feature inconsistency by explicitly increasing the amount of consistent (i.e., fused) features in the fusion process. Instead of completely fusing all the features extracted from images of different spectra, we apply an alternative scheme where at each convolution level, only one part of thermal and visible features (respectively represented by green and yellow blocks in Figure 4.2 right) are fused, and the proportion of the fused features (represented by purple blocks) gradually increases throughout multiple convolution levels. More concretely, for a given convolution level, monospectral features are generated via asymmetric fusions while fused features are generated via symmetric fusions. Figure 4.3 summarizes their detailed structures.

4.2.1 Asymmetric fusions

As illustrated in Figure 4.3 left, to generate the thermal features (subscript t) of the i^{th} convolution level (f_t^i), the thermal features and the fused (subscript f) features of the previous levels (f_t^{i-1} and f_f^{i-1}) are used. This fusion can be formalized as:

$$f_t^i = \mathcal{F}_1(\sigma(f_t^{i-1}, f_f^{i-1})) \tag{4.6}$$

where σ is a feature concatenation operation; $\mathcal{F}_1$ is a 3×3 convolution operation followed by a BN layer and a ReLU function. To extract visible features (subscript v) of level i (f_v^i), we have similarly:

$$f_v^i = \mathcal{F}_2(\sigma(f_v^{i-1}, f_f^{i-1})) \tag{4.7}$$

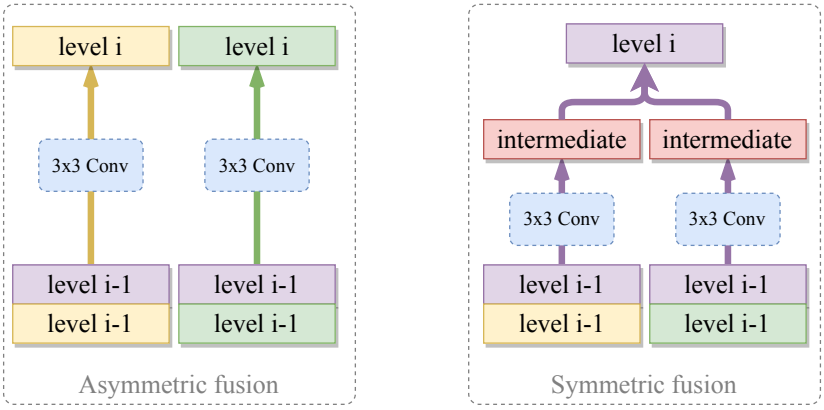


FIGURE 4.3 – The proposed asymmetric and symmetric fusions in PS-Fuse. Yellow, green and purple blocks represent thermal, visible and fused features. Better viewed in colour.

4.2.2 Symmetric fusions

As illustrated in Figure 4.3 right, the fused features (f_f^i) are generated using the fused features (f_f^{i-1}) and both monospectral features (f_t^{i-1} and f_v^{i-1}):

$$f_f^i = \psi(\mathcal{F}_3(\sigma(f_t^{i-1}, f_f^{i-1})), \mathcal{F}_4(\sigma(f_v^{i-1}, f_f^{i-1}))) \tag{4.8}$$

where σ is the feature concatenation operation; $\mathcal{F}_3$ and $\mathcal{F}_4$ are both 3×3 convolution layers with BN; ψ is the element-wise average operation followed by a ReLU function for feature fusion.

4.3 Experimental results for CFR and PS-Fuse

Network architecture. We implemented our Cyclic Fuse-and-Refine (CFR) and Progressive Spectral Fusion (PS-Fuse) on the single stage object detector FSSD (Z. Li & Zhou, 2017), which is an improved version of the well-known SSD detector (W. Liu et al., 2016). Following previous works, the monospectral features are extracted independently through a VGG-16 (Simonyan & Zisserman, 2015) network, and fused after the conv4_3 layer (halfway through the network). To ensure a fair comparison, the element-wise average operation is used as the baseline fusion method. For CFR, we integrate the proposed module with a different number of loops (1-4). For PS-Fuse, conv4_1, conv4_2 and conv4_3 layers are involved for the progressive fusion. Specifically, two experimental settings are proposed, whether the number of monospectral features is halved after a convolution level (Setting A), or the number of fused features is doubled after a convolution level (Setting B).

Comparison with state-of-the-art methods. In Table 4.1, we compare the experimental results of our approach with state-of-the-art methods on KAIST dataset (Hwang et al., 2015). For these experiments, we make 3 loops in the Fuse-and-Refine cycle, and choose Setting A of PS-Fuse as previously described. Depending

Methods	Miss Rate (lower, better)		
	All	Day	Night
Training with original annotations:			
ACF+T+THOG (Hwang et al., 2015)	47.24%	42.44%	56.17%
Halfway Fusion (J. Liu et al., 2016)	26.15%	24.85%	27.59%
Fusion RPN+BF (Konig et al., 2017)	16.53%	16.39%	18.16%
IAF R-CNN (C. Li et al., 2019)	16.22%	13.94%	18.28%
IATDNN+IASS (Guan et al., 2019)	15.78%	15.08%	17.22%
MSDS-RCNN (C. Li et al., 2018)	11.63%	10.60%	13.73%
CFR_3	**10.05%**	**9.72%**	10.80%
PS-Fuse (A)	10.07%	10.73%	**8.96%**
Training with sanitized annotations:			
MSDS-RCNN (C. Li et al., 2018)	7.49%	8.09%	5.92%
CFR_3	**6.13%**	**7.68%**	3.19%
PS-Fuse (A)	6.35%	8.51%	**2.46%**

TABLE 4.1 – Detection accuracy comparisons in terms of Miss Rate percentage on KAIST dataset (Hwang et al., 2015). Our competitors' results are taken from (C. Li et al., 2018).

Methods	Bicycle	Car	Person	mAP
Baseline	56.39%	83.90%	73.28%	71.17%
CFR_3	**57.77%**	**84.91%**	74.49%	**72.39%**
PS-Fuse (A)	57.15%	84.14%	**75.00%**	72.10%

TABLE 4.2 – Detection accuracy comparisons in terms of mean Average Precision on FLIR dataset.

on what was done in the literature and to allow a fair comparison, we report our detection accuracy with original and "sanitized" training annotations, respectively. All the compared deep learning-based methods use the same input image resolution (640×512) and the same backbone network (VGG-16). The results show that our proposed methods allow us to obtain better detection results than all their competitors for both training annotations. In Table 4.2, we compare the mAP of three different models: a baseline model which uses the traditional halfway fusion architecture and our proposed CFR and PS-Fuse. Again, our method provides important precision gains for all the considered object categories.

Ablation study. We study in details the effectiveness of the proposed CFR module and the relationship between the number of loops in the fuse-and-refine cycle and the reduction of multispectral feature inconsistency. The experimental results are summarized in Table 4.3. We provide the Miss Rate and the L2 distance between thermal/visible features before/after each refinement. These distances are used as an indicator of the similarity between thermal and visible features. From the table we observe successive accuracy gains from 1 to 3 loops, and a decrease after 4 loops;

Methods	Miss rates	L2 distances $\times 100$	Param.	FLOPs
CFR (1)	6.90%	$\{2.04, 1.96\}$	16.53M	66.52B
CFR (2)	6.40%	$\{2.04, 1.96, 1.86\}$	16.53M	72.56B
CFR (3)	**6.13%**	$\{1.94, 1.85, 1.76, 1.66\}$	16.53M	78.60B
CFR (4)	7.09%	$\{1.88, 1.78, 1.71, 1.63, 1.55\}$	16.53M	84.64B

TABLE 4.3 – Miss rates versus L2 distances with respect to different numbers of Fuse-and-Refine loops. We also report the number of parameters and FLOPs.

Methods	Miss Rate (lower, better)			Param.	FLOPs
	R-All	R-Day	R-Night		
Baseline	7.68%	10.05%	3.40%	11.81M	60.48B
PS-Fuse (A)	**6.35%**	**8.51%**	**2.46%**	9.15M	46.83B
PS-Fuse (B)	6.80%	8.94%	2.70%	10.33M	52.89B

TABLE 4.4 – Miss rates for different PS-Fuse architectures.

Methods	Miss Rate (lower, better)		
	All	Day	Night
Thermal only	20.63%	24.46%	11.27%
Visible only	24.92%	18.18%	38.87%
Baseline multispectral	7.68%	10.05%	3.40%
CFR thermal branch	7.29%	8.14%	5.86%
CFR visible branch	7.71%	8.75%	5.64%
CFR multispectral	**6.13%**	**7.68%**	3.19%
PS-Fuse thermal branch	8.13%	8.21%	7.72%
PS-Fuse visible branch	8.98%	9.53%	7.85%
PS-Fuse multispectral	6.35%	8.51%	**2.46%**

TABLE 4.5 – Monospectral and multispectral results on KAIST dataset.

meanwhile, the value of L2 distance continues to decrease along with the number of loops. As mentioned in Section 2.2, the lack of consistency between the multispectral features is harmful; on the contrary, if they are too consistent, we notice sharp emerge/plunge in the feature values which makes the fusion meaningless. This explains why the Miss Rate starts to increase after 4 loops. We compare in Table 4.4 the performance of different PS-Fuse architectures. The baseline halfway fusion is the traditional complete feature fusion after the conv4_3 layer of VGG-16. Then, we replace the baseline halfway fusion by the proposed PS-Fuse. We implemented the settings A and B as previous explained. From the table, we can observe some accuracy improvements for both settings (1.33% and 0.88% of improvements from Setting A and B).

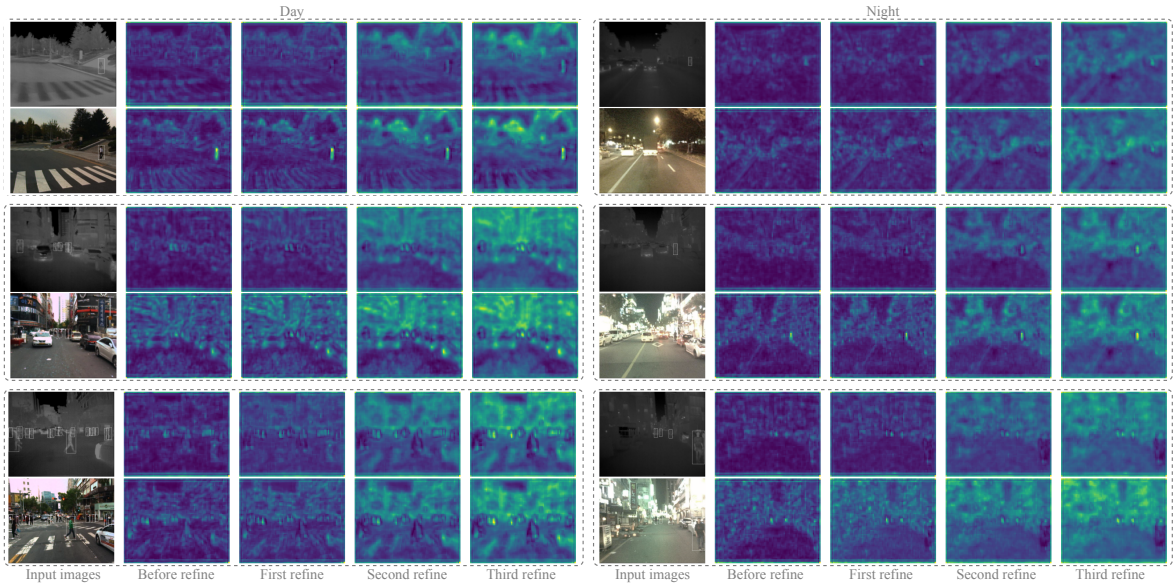


FIGURE 4.4 – Examples of thermal/visible image pairs and their corresponding features before/after fuse-and-refine operations in Cyclic Fuse-and-Refine (CFR). These multispectral image pairs are taken from KAIST for the pedestrian detection task. Zoom in to see details.

Discussion. In order to verify the validity of the underlying principle for reducing multispectral feature inconsistency, we have conducted some additional experiments to evaluate the monospectral performance of the baseline fusion methods against CFR and PS-Fuse. Our results are listed in Table 4.5: "Thermal only" and "Visible only" represent the "pure" monospectral results (equivalent to standard object detector); "Baseline multispectral" is the normal halfway fusion of thermal and visible features (baseline results). From these results, we can hypothesize that the thermal features and visible features are inconsistent since the "day", "night" and "all" Miss Rate performance are very different (6.28%, 27.6%, and 4.29% of absolute difference, respectively).

We then report the results of CFR thermal features, visible features and fused features. The results of CFR indicate the reduction of multispectral inconsistency (less Miss Rate difference between thermal and visible) and the improvement of the monospectral features' quality (lower Miss Rate than "Thermal only" or "Visible only"). To visualize this process, we show in Figure 4.4 some examples of multispectral images and their corresponding features before/after fuse-and-refine operations. It can be observed that the features of one modality are gradually corrected by introducing the information from the other modality, e.g., on the first line we show two examples (day & night) where the pedestrian is barely visible from one modality but features from the other modality help to successfully localize the "invisible" pedestrian.

We also report the results of PS-Fuse thermal features, visible features and fused features. It can be observed that after integrating the fused features into the monospectral ones, thermal features and visible features become more consistent and more relevant (less Miss Rate difference between "PS-Fuse thermal branch" and "PS-Fuse

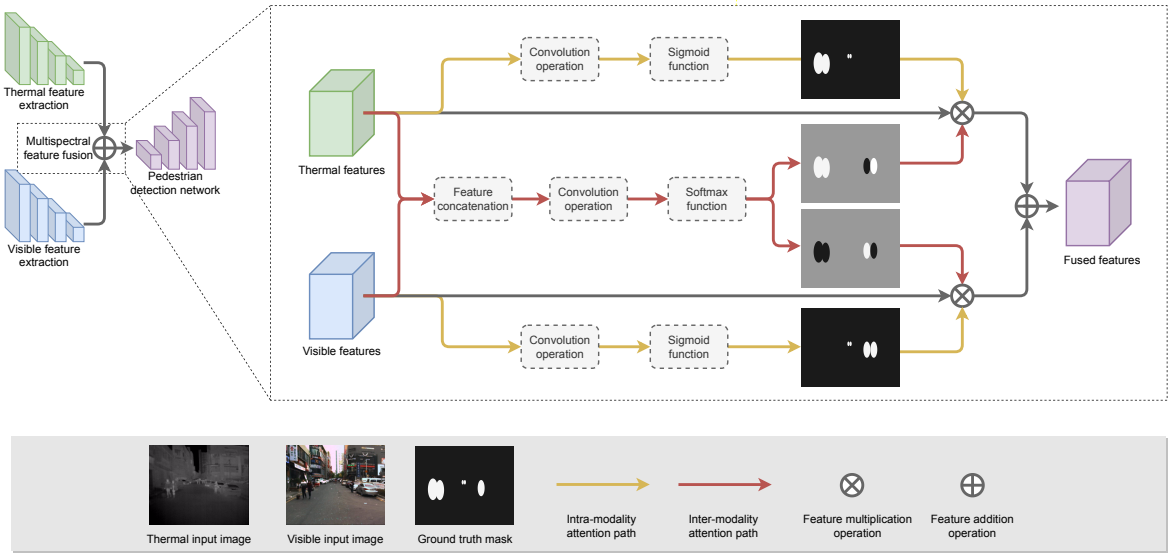


FIGURE 4.5 – The overall architecture of GAFF. Green, blue and purple blocks represent thermal, visible and fused features. Yellow and red paths represent the intra- and inter-modality attention modules.

visible branch", lower Miss Rate than "Thermal only" and "Visible only"). Consequently, their combination "PS-Fuse multispectral" gets better results than "Baseline multispectral".

Another remarkable result from Table 4.5 is that, regardless of the fusion methods, multispectral results are always better than monospectral results, e.g., even if the thermal channel is more informative than the visible one at nighttime, multispectral nighttime Miss Rate still outperforms its thermal counterpart. This is because in KAIST and FLIR datasets, the images are acquired in an urban environment, where the lighting is sufficient at nighttime. Therefore, both thermal and visible modalities are informative in most cases.

4.4 Guided Attentive Feature Fusion

In our third proposed method, **Guided Attentive Feature Fusion** (GAFF), we seek to identify the more reliable modality when the two modalities produce contradictory representations. An intuitive solution of this automatic identification is to manually set multiple usage scenarios and design a specific solution for each scenario. For example, (Guan et al., 2019) proposes an illumination-aware network consisting of a day illumination subnetwork and a night illumination subnetwork. The detection results from the two subnetworks are then fused according to the prediction of the illumination value. Such kind of hand-crafted fusion mechanism improves the resilience of the model to a certain extent, nonetheless, there are still two limitations: Firstly, cherry-picked scenarios may not cover all the conditions, e.g., different illumination/season/weather conditions; Secondly, the situation may be completely different even in the same usage scenario, e.g., at nighttime, lighting conditions in urban areas are different from those in rural areas.

The proposed GAFF method is, however, a fully adaptive approach. By combining the intra-modality and the inter-modality attention modules, GAFF allows the network to learn the adaptive weighing and fusion of multispectral features. These two attention mechanisms are guided by the prediction and comparison of the pedestrian masks in the multispectral feature fusion stage. Specifically, at each spatial position, thermal or visible features are enhanced when they are located in the area of a pedestrian (intra-modality attention) or when they possess a higher quality than in the other modality (inter-modality attention). More implementation details on these two attention modules are given below.

4.4.1 Intra-modality attention module

The intra-modality attention module aims at enhancing the thermal or visible features in a monospectral view. Specifically, as illustrated by the yellow paths on Figure 4.5, features of an area with a pedestrian are highlighted by multiplying the learnt features with the predicted pedestrian mask. Moreover, in order to avoid directly affecting the thermal or visible features, the highlighted features are added as a residual to enhance the monospectral features. It can be formalized as:

$$\begin{aligned} f^t_{intra} &= f^t \otimes (1 + m^t_{intra}) \\ f^v_{intra} &= f^v \otimes (1 + m^v_{intra}) \end{aligned} \tag{4.9}$$

where

$$\begin{aligned} m^t_{intra} &= \sigma(\mathcal{F}^t_{intra}(f^t)) \\ m^v_{intra} &= \sigma(\mathcal{F}^v_{intra}(f^v)) \end{aligned} \tag{4.10}$$

Superscripts (t or v) denote the thermal (t) or visible (v) modality; $\otimes$ denotes the element-wise multiplication operation; σ represents the Sigmoid function[1]; $\mathcal{F}_{intra}$ represents a convolution operation to predict the intra-modality attention masks (pedestrian masks) m_{intra}; f and f_{intra} represent the original and enhanced features, respectively.

The prediction of the pedestrian mask is supervised by the semantic segmentation loss, where the ground truth mask (m^{gt}_{intra}) is converted from the object detection annotations. As illustrated in Figure 4.5, the bounding box annotations are transformed into filled ellipses to approximate the shape of the true pedestrians.

4.4.2 Inter-modality attention module

Thermal and visible cameras have their own imaging characteristics, and under certain conditions, one sensor has superior imaging quality (i.e. is more relevant for the considered task) than the other. To leverage both modalities, we propose the

[1]Sigmoid function is defined as: $S(x) = \frac{1}{1+e^{-x}}$

inter-modality attention module, which adaptively selects thermal or visible features according to the dynamic comparison of their feature quality. Concretely, an inter-modality attention mask is predicted based on the combination of thermal and visible features. This predicted mask has two values for each pixel, corresponding to the weights for thermal and visible features (summing to 1). This attention module is illustrated as the red paths in Figure 4.5. It can be formulated as:

$$\begin{aligned} f^t_{inter} &= f^t \otimes (1 + m^t_{inter}) \\ f^v_{inter} &= f^v \otimes (1 + m^v_{inter}) \end{aligned} \tag{4.11}$$

where

$$m^t_{inter}, m^v_{inter} = \delta(\mathcal{F}_{inter}([f^t, f^v])) \tag{4.12}$$

Here, δ denotes the Softmax function[2]; $[\cdot]$ denotes the feature concatenation operation; $\mathcal{F}_{inter}$ represents a convolution operation to predict the inter-modality attention mask m_{inter}. At each spatial position of the mask, the sum of m^t_{inter} and m^v_{inter} equals to 1. Following the same principles, this formalization could theoretically allow for more than two modalities to be fused.

The inter-modality attention module allows the network to adaptively select the most reliable modality. However, in order to train this module, we should need a pixel-level ground truth information about the best modality quality. Our solution to relieve the annotation cost is to assign labels according to the prediction error of the pedestrian masks from the intra-modality attention module, i.e., we force the network to select one modality if its intra-modality mask prediction is better (i.e. closer to the ground truth pedestrian mask) than the other. Specifically, we first calculate an error mask for each spectrum with the following formula:

$$\begin{aligned} e^t_{intra} &= \mid m^t_{intra} - m^{gt}_{intra} \mid \\ e^v_{intra} &= \mid m^v_{intra} - m^{gt}_{intra} \mid \end{aligned} \tag{4.13}$$

then the label for the modality selection is defined as:

$$m^{gt}_{inter} = \begin{cases} 1,0 & if\ (e^v_{intra} - e^t_{intra}) > margin \\ 0,1 & if\ (e^t_{intra} - e^v_{intra}) > margin \\ ignored & otherwise \end{cases} \tag{4.14}$$

Here, $|\cdot|$ denotes the absolute function; e_{intra} represents the error mask, defined by the difference between the predicted intra-modality mask m_{intra} and the ground truth intra-modality mask m^{gt}_{intra}; m^{gt}_{inter} is the ground truth mask for inter-modality attention (2 values at each mask position); *margin* is a hyperparameter to be tuned.

[2]Softmax function is defined as: $\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \quad for\ i = 1, 2, \ldots, K$

An example of the label assignment for the inter-modality attention mask is shown in Figure 4.5. If the intra-modality pedestrian masks are predicted as shown in the yellow paths, the inter-modality (weak) ground truth masks are then defined as the ones shown on the red paths, where white, black and grey areas denote the classification labels 1, 0 and ignored, respectively. Here, the thermal features produce a better intra-modality mask prediction for the pedestrians on the left side of the input images in Figure 4.5. Therefore, according to Equation 4.14, the label for the inter-modality mask on this area is assigned as 1, 0 (1 for the thermal mask and 0 for the visible mask). For regions where the two intra-modality masks have comparable prediction qualities (i.e., the difference between prediction errors is smaller than the predefined margin), the optimization of the inter-modality attention mask prediction on these areas are ignored (i.e., do not participate in the loss calculation).

4.4.3 Combining intra- and inter-modality attention

The intra-modality attention module enhances features on areas with pedestrians, and the inter-modality attention module adaptively selects features from the most reliable modality. When these two modules are combined, the fused features are obtained by:

$$f^{fused} = \frac{f^t_{hybrid} + f^v_{hybrid}}{2} \tag{4.15}$$

where

$$\begin{aligned} f^t_{hybrid} &= f^t \otimes (1 + m^t_{intra}) \otimes (1 + m^t_{inter}) \\ f^v_{hybrid} &= f^v \otimes (1 + m^v_{intra}) \otimes (1 + m^v_{inter}) \end{aligned} \tag{4.16}$$

Here, m_{intra} and m_{inter} are predicted intra- and inter-modality attention masks from Equation 4.10 and Equation 4.12; f_{hybrid} represents features enhanced by both attention modules; f^{fused} represents the final fused features.

As mentioned in Section 2.2, the optimization of the multispectral feature fusion task may not benefit enough from the sole optimization of the object detection task (as done e.g. in (L. Zhang, Liu, Zhang, et al., 2019)). In GAFF, we propose two specific feature fusion losses, including the pedestrian segmentation loss for the intra-modality attention and the modality selection loss for the inter-modality attention, to guide the multispectral feature fusion task. These losses are jointly optimized with the object detection loss. The final training loss $\mathcal{L}_{total}$ is calculated as:

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \mathcal{L}_{intra} + \mathcal{L}_{inter} \tag{4.17}$$

where, $\mathcal{L}_{det}$, $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ are the pedestrian detection, the intra- and inter-modality attention loss, respectively.

Margin	Miss Rate		
	All	Day	Night
0.05	6.92%	8.47%	3.68%
0.1	**6.48%**	**8.35%**	**3.46%**
0.2	7.47%	9.31%	4.22%

TABLE 4.6 – Detection results of GAFF with different *margin* values in the inter-modality attention module.

Residual	Miss Rate		
	All	Day	Night
	7.46%	8.88%	4.85%
✓	**6.48%**	**8.35%**	**3.46%**

TABLE 4.7 – Detection results of GAFF where the attention masks are directly applied or added as residual.

4.5 Experimental results for GAFF

Implementation details. The proposed GAFF module can be included in any type of two-stream convolutional neural networks. In these experiments, we choose RetinaNet (Lin, Goyal, et al., 2017) as our base detector. It is transformed into a two-stream convolutional neural network by adding a backbone branch for the extraction of thermal features. A ResNet-18 (He et al., 2016) or a VGG-16 (Simonyan & Zisserman, 2015) network is pre-trained on ImageNet (Deng et al., 2009), then adopted as our backbone network. The input image resolution is fixed to 640×512 for training and evaluation. Our baseline detector applies the basic addition operation as the multispectral feature fusion method. GAFF is implemented by adding the intra- and inter-modality attention modules, corresponding to the yellow and the red branches in Figure 4.5. Focal loss (Lin, Goyal, et al., 2017) and Balanced L1 loss (Pang et al., 2019) are adopted as the classification loss and the bounding box regression loss to optimize the object detection task. In order to introduce our specific guidance, we adopt the DICE (Dice, 1945) loss as the pedestrian segmentation loss ($\mathcal{L}_{intra}$ in Equation 4.17) and the cross-entropy loss as the modality selection loss ($\mathcal{L}_{inter}$ in Equation 4.17).

Hyperparameter tuning. As reported in Table 4.6, we conduct experiments with different *margin* values in the inter-modality attention module on KAIST dataset (Hwang et al., 2015) with "sanitized" annotations. The Miss Rate scores on the Reasonable-all, Reasonable-day and Reasonable-night subsets are listed. We observe that the optimal Miss Rate is achieved when $margin = 0.1$. Thus, we use $margin = 0.1$ for all the following experiments.

Residual attention. As mentioned earlier, attention enhanced features are added as residual to avoid directly affecting the thermal or visible features. We verify this

Backbone	GAFF		Miss Rate		
	Intra.	Inter.	All	Day	Night
ResNet-18			13.04%	13.83%	11.60%
	✓		12.13%	11.97%	11.99%
		✓	11.15%	10.68%	11.67%
	✓	✓	**10.74%**	**10.46%**	**11.10%**
VGG-16			12.72%	11.37%	15.57%
	✓		11.78%	11.45%	12.50%
		✓	11.03%	10.99%	11.44%
	✓	✓	**10.62%**	**10.82%**	**10.14%**

TABLE 4.8 – Ablation study of two attentive fusion modules on KAIST dataset (Hwang et al., 2015) with original annotations.

Backbone	GAFF		Miss Rate		
	Intra.	Inter.	All	Day	Night
ResNet-18			9.98%	12.46%	5.29%
	✓		9.26%	11.51%	5.32%
		✓	9.29%	11.97%	5.14%
	✓	✓	**7.93%**	**9.79%**	**4.33%**
VGG-16			9.28%	11.73%	5.17%
	✓		8.70%	11.42%	3.55%
		✓	7.73%	10.35%	**2.81%**
	✓	✓	**6.48%**	**8.35%**	3.46%

TABLE 4.9 – Ablation study of two attentive fusion modules on KAIST dataset (Hwang et al., 2015) with "sanitized" annotations.

choice by comparing in Table 4.7 the Miss Rate of GAFF where the attention masks are directly applied to monospectral features ($f_{intra} = f \otimes m_{intra}$ and $f_{inter} = f \otimes m_{inter}$) or added as residual (as in Equation 4.9 and Equation 4.11).

Necessity of attention. We compare in Table 4.8 and 4.9 the detection accuracy on KAIST dataset with different attention settings, different backbone networks, and different annotation settings (original and "sanitized"). When conducting experiments with inter-modality but without intra-modality attention, the pedestrian masks are predicted but are not multiplied with the corresponding monospectral features. For each backbone network or annotation setting, both intra- and inter-modality attention modules consistently improve the baseline detection accuracy, and their combination leads to the lowest overall Miss Rate under all experimental settings. The present findings confirm the effectiveness of the proposed guided attentive feature fusion modules.

Backbone	Guidance	Miss Rate		
		All	Day	Night
ResNet-18		13.15%	13.71%	11.54%
	✓	**10.74%**	**10.46%**	**11.10%**
VGG-16		13.67%	13.19%	14.51%
	✓	**10.62%**	**10.82%**	**10.14%**

TABLE 4.10 – Comparison between guided and non-guided models on KAIST dataset with original annotations.

Backbone	Guidance	Miss Rate		
		All	Day	Night
ResNet-18		9.05%	10.63%	6.01%
	✓	**7.93%**	**9.79%**	**4.33%**
VGG-16		8.38%	10.39%	4.44%
	✓	**6.48%**	**8.35%**	**3.46%**

TABLE 4.11 – Comparison between guided and non-guided models on KAIST dataset with "sanitized" annotations.

Necessity of guidance. To explore the effects of the proposed multispectral feature fusion guidance, we compare our guided approach to one with a similar network architecture as ours but where the optimization of the specific fusion losses ($\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$ in Equation 4.17) are removed from the training process, i.e., the fusion is only supervised by the object detection loss (as done with (L. Zhang, Liu, Zhang, et al., 2019)). We report in Table 4.10 and 4.11 the detection performance with and without guidance, under different backbone networks and annotations settings. The results confirm our assumption that the object detection loss is not relevant enough for the multispectral feature fusion task: even though the non-guided attentive fusion module improves the baseline Miss Rate to some degree (e.g., with the "sanitized" annotations and VGG-16 backbone, non-guided model improves the base detector's Miss Rate from 9.28% to 8.38%), it could be further improved when the specific fusion guidance is added (from 8.38% to 6.48%).

Attention mask interpretation. Figure 4.6 and 4.7 provides the visualization results of the intra-modality, the inter-modality and the hybrid attention masks during daytime and nighttime. For each figure, the top and bottom two rows of images are visualization results of guided and non-guided attentive feature fusions, respectively. We can see on the intra-modality attention masks that the guided attention mechanism focuses on pedestrian areas, even though, sometimes, it is not accurate from a single monospectral view. For example, the traffic cone is misclassified as a pedestrian due to its human-like shape on the thermal image of Figure 4.6, and the pedestrian in the middle right position is missed due to insufficient lighting on the RGB image of Figure 4.7. For inter-modality attention masks, it appears that the guided attentive fusion tends to select visible features on well-lit areas (such as

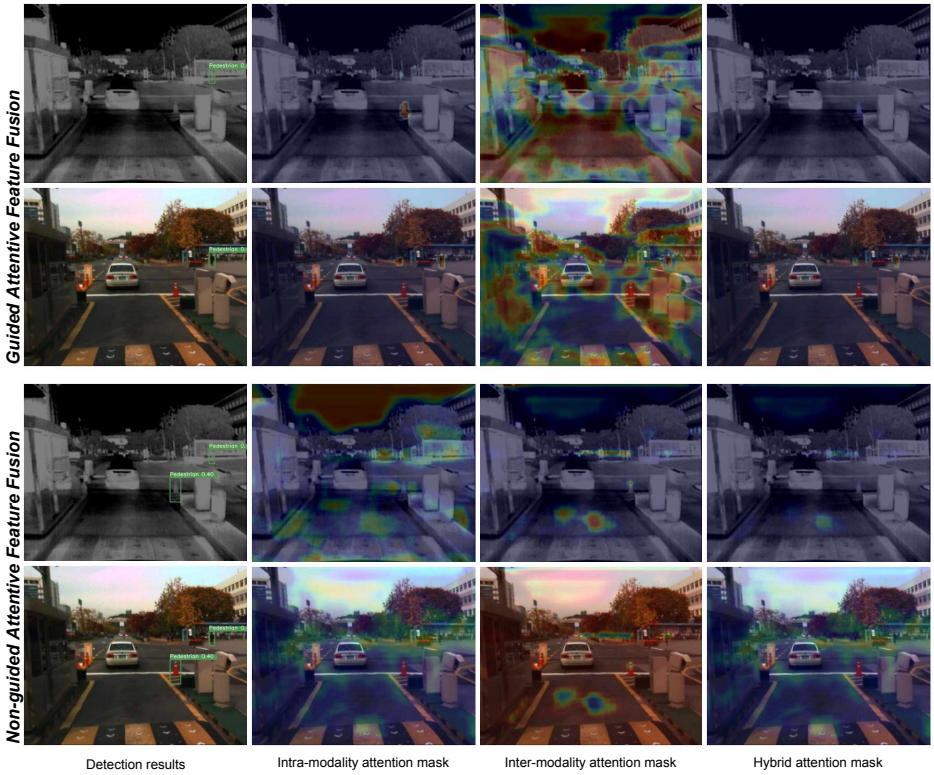


FIGURE 4.6 – Visualization examples of attention masks during daytime on KAIST dataset.

upside of images in Figure 4.7) and brightly coloured areas (e.g., traffic cone, road sign, speed bump, car tail light, etc), and to select thermal features on dark areas and uniform areas (such as sky and road). Note that these attention preferences are automatically learnt via our inter-modality attention guidance. On the contrary, despite the fact that the non-guided attention mechanism brings some accuracy improvements, the predicted attention masks are quite difficult to interpret. More visualization results are shown in Figure 4.8 and 4.9. Besides, an interesting error case is shown in Figure 4.10, where the pedestrian on the steps is not detected with the guided model but detected with the non-guided model. As mentioned earlier, GAFF selects thermal features on uniform areas, which is intuitive since thermal cameras are sensitive to temperature change and there exist few objects on uniform areas of the thermal image. However, in this particular case, the pedestrian is not captured on the thermal image, which leads to the final detection error.

Attention accuracy evolution. We plot in Figure 4.11 the evolution of intra- and inter-modality attention accuracy during training. Specifically, red solid and dashed lines represent the pedestrian segmentation accuracy (via DICE score (Dice, 1945) $Dice = \frac{2|A \cap B|}{|A|+|B|}$) from thermal and visible features in intra-modality attention module; blue line indicates the modality selection accuracy in inter-modality attention module. From the plot, we can conclude that thermal images are generally better for recognition than RGB images. This observation is consistent with our monospectral

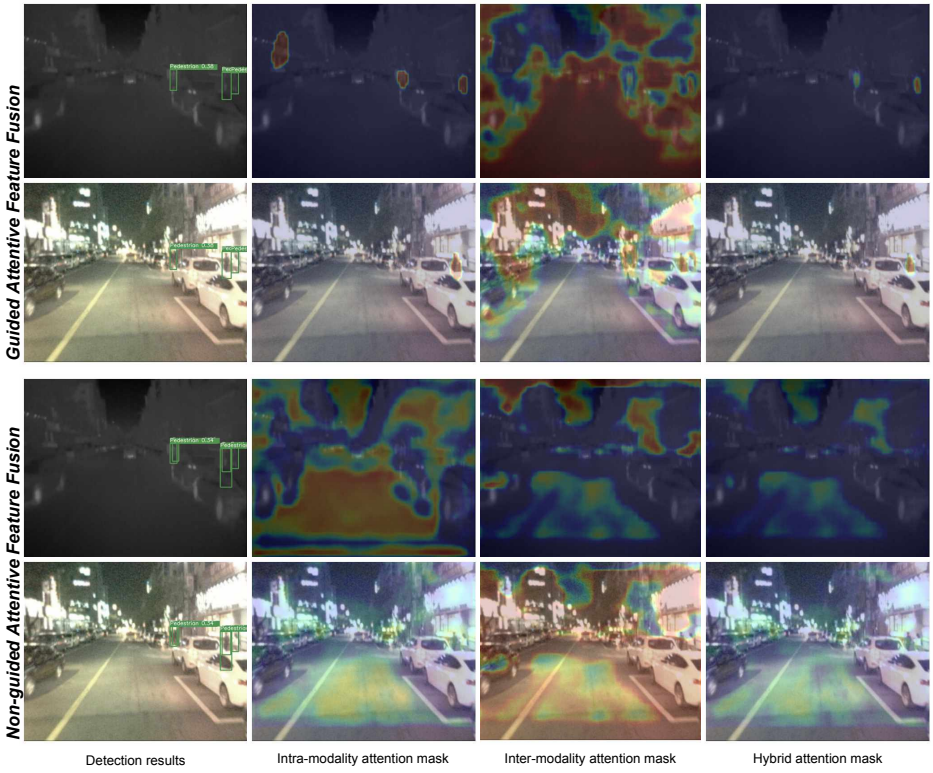


FIGURE 4.7 – Visualization examples of attention masks during nighttime on KAIST dataset.

Backbone	GAFF	Param.	Runtime	
			1080Ti	TX2
ResNet-18		23,751,725	10.31ms	10.5ms
	✓	23,765,553	10.85ms	12.1ms
VGG-16		31,403,053	8.87ms	10.3ms
	✓	31,430,705	9.34ms	11.6ms

TABLE 4.12 – Runtime on different computing platforms.

experiments, where thermal-only model reaches 18.8% of Miss Rate while visible-only model achieves 20.74% (both trained with "sanitized" annotations). Interestingly, as the segmentation accuracy increases for both images, the modality selection task becomes more and more challenging. Note that this accuracy is irrelevant at the beginning of the training, where predicted pedestrian masks are almost zero for both thermal and visible features, thus the difference between their error masks is minor and the set of *margin* makes most areas ignored for modality selection optimization. Such a mechanism avoids the "cold start" problem.

Runtime analysis. In Table 4.12 we report the total number of learnable parameters and the average inference runtime on two different computation platforms. Specifically, the models are implemented with Pytorch (TensorRT) framework for an inference time testing on the Nvidia GTX 1080Ti (Nvidia TX2) platform. Since GAFF

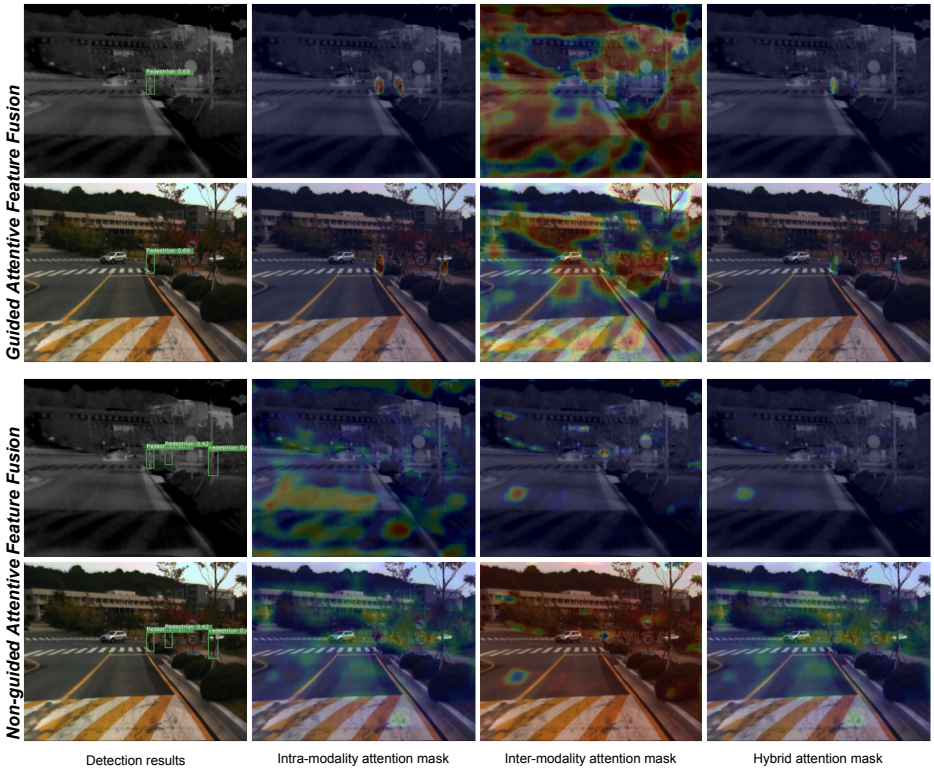


FIGURE 4.8 – More visualization examples of attention masks during daytime on KAIST dataset.

only involves 3 convolution layers, the additional parameters and computation cost is low, i.e., it represents less than 0.1% of additional parameters and around 0.5ms (1.5ms) of inference time on 1080Ti (TX2). Note that the time for post-processing treatments (such as Non-Maximum Suppression) is not taken into account for the benchmarking. Our model meets the requirement of real-time treatment on embedded devices, which is essential for many applications.

Comparison with State-of-the-Art methods. Table 4.13 shows the detection results of existing methods and our GAFF with the original and "sanitized" annotations on KAIST. It can be observed that GAFF achieves state-of-the-art performance on this dataset. According to Table 4.14, thanks to the lightweight design of GAFF, our model has substantial advantage in terms of inference speed.

Table 4.15 reports the detection results with and without GAFF on FLIR dataset. We can observe that the average precision is improved for all IoU thresholds with GAFF (around 1% of mAP improvement for both backbone networks), which shows that our method can generalize well to different types of images.

Methods	Miss Rate		
	All	Day	Night
Training with original annotations:			
ACF+T+THOG (Hwang et al., 2015)	47.24%	42.44%	56.17%
Halfway Fusion (Konig et al., 2017)	26.15%	24.85%	27.59%
Fusion RPN+BF (Konig et al., 2017)	16.53%	16.39%	18.16%
IAF R-CNN (C. Li et al., 2019)	16.22%	13.94%	18.28%
IATDNN+IASS (Guan et al., 2019)	15.78%	15.08%	17.22%
CIAN (L. Zhang, Liu, Zhang, et al., 2019)	14.12%	14.77%	11.13%
MSDS-RCNN (C. Li et al., 2018)	11.63%	**10.60%**	13.73%
GAFF (ours)	**10.62%**	10.82%	**10.14%**
Training with sanitized annotations:			
MSDS-RCNN (C. Li et al., 2018)	7.49%	**8.09%**	5.92%
GAFF (ours)	**6.48%**	8.35%	**3.46%**

TABLE 4.13 – Detection results on KAIST dataset.

Methods	Platform	Runtime
ACF+T+THOG (Hwang et al., 2015)	MATLAB	2730ms
Halfway Fusion (Konig et al., 2017)	Titan X	430ms
Fusion RPN+BF (Konig et al., 2017)	MATLAB	800ms
IAF R-CNN (C. Li et al., 2019)	Titan X	210ms
IATDNN+IASS (Guan et al., 2019)	Titan X	250ms
CIAN (L. Zhang, Liu, Zhang, et al., 2019)	1080Ti	70ms
MSDS-RCNN (C. Li et al., 2018)	Titan X	220ms
GAFF (ours)	1080Ti	**9.34ms**

TABLE 4.14 – Runtime comparisons on KAIST dataset.

Backbone	GAFF	mAP	AP75	AP50
ResNet-18		36.6%	31.9%	72.8%
	✓	**37.5%**	**32.9%**	**72.9%**
VGG-16		36.3%	30.2%	71.9%
	✓	**37.3%**	**30.9%**	**72.7%**

TABLE 4.15 – Detection results on FLIR dataset.

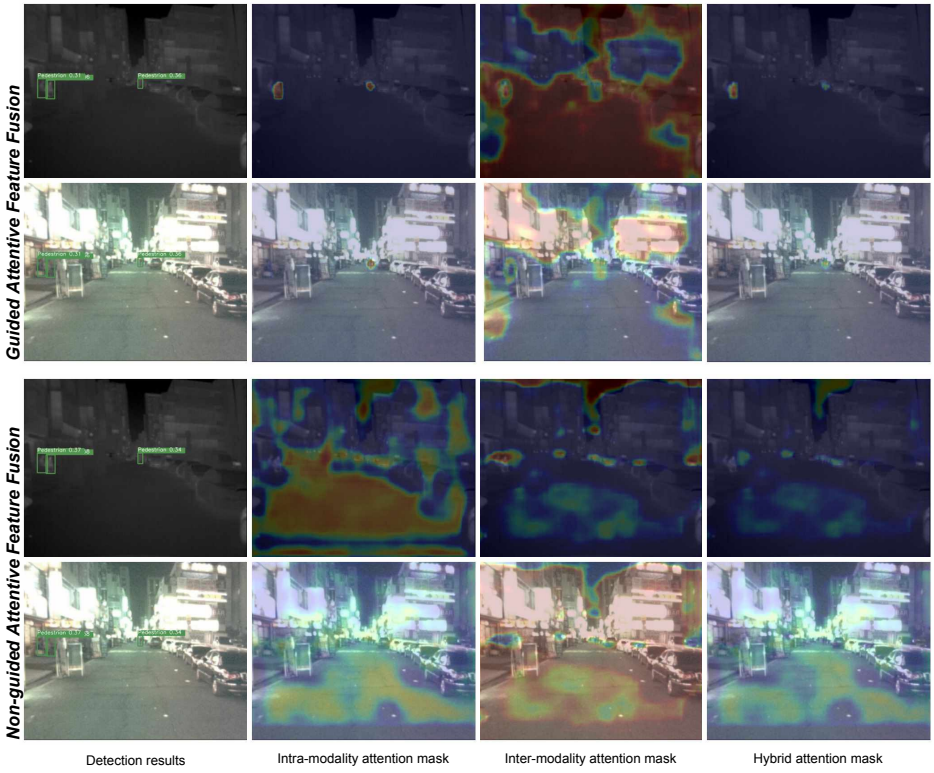


FIGURE 4.9 – More visualization examples of attention masks during nighttime on KAIST dataset.

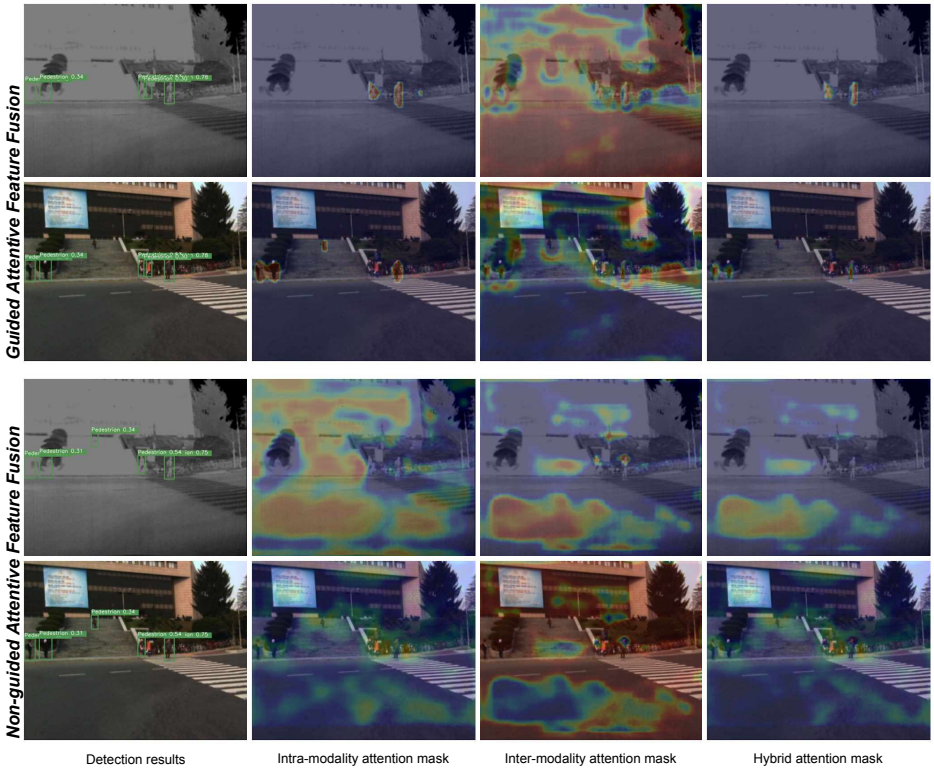


FIGURE 4.10 – Error cases of attention masks.

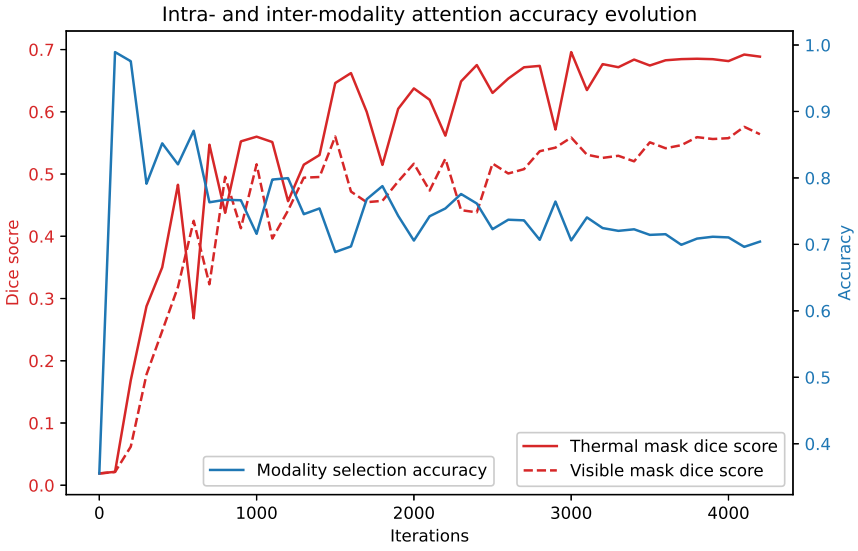
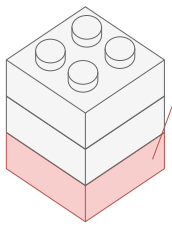


FIGURE 4.11 – Intra- and inter-modality attention accuracy evolution during training.

Chapter 5

Sensors and annotations: low cost multispectral data processing

Contents



Data collection

This chapter introduces our methods to reduce multispectral sensor cost and human annotation efforts.

As is known, high resolution and manually annotated data are essential for multispectral scene analysis via supervised learning. However, in the actual product development process, the manufacture and labour costs are factors that have to be considered. Therefore, we apply Active Learning (AL) and Knowledge Distillation (KD) techniques to reduce the aforementioned costs, while minimizing the performance degradation. We hope that these practical approaches can be used in actual industrial developments.

This chapter concerns the following publications:

"Deep Active Learning from Multispectral Data Through Cross-Modality Prediction Inconsistency" in *28th International Conference on Image Processing (ICIP2021)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

"Low Cost Multispectral Scene Analysis with Modality Distillation", in *Winter Conference on Applications of Computer Vision (WACV2022)*
Heng Zhang, Elisa FROMONT, Sébastien LEFEVRE, Bruno AVIGNON

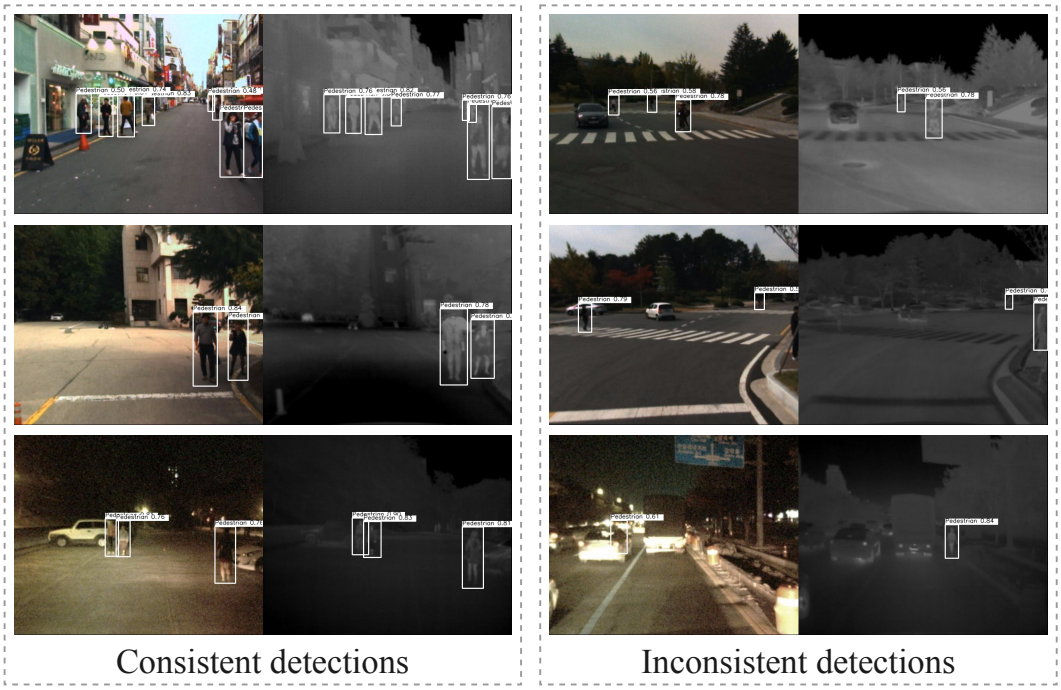


FIGURE 5.1 – Example of multispectral image pairs and their corresponding monospectral pedestrian detection results.

5.1 Deep Active Learning from Multispectral Data

As explained in Section 2.4, collecting labelled multispectral data is expensive and time-consuming, which motivates us to build an accurate multispectral scene analysis system with minimal annotation efforts via an Active Learning (AL) strategy.

In Figure 5.1, we show some image pairs from visible & thermal cameras of identical scenes and their corresponding monospectral pedestrian detection results. Note that the image acquisition and the pedestrian detection from the two modalities are completely independent. We split these multispectral image pairs into two categories: pairs with consistent detections (on the left side of Figure 5.1) and inconsistent detections (on the right side). From these image pairs, we can observe that the detection results from the two modalities are similar in most cases, which indicates the **redundancy** for a multispectral system; whereas at least one modality is wrong when the detections are contradictory, which demonstrates the **complementarity** of multispectral systems.

We suggest relying on the **complementarity** of different sensors for the adaptive selection of multispectral samples to be annotated. Specifically, our proposed active criterion is based on the **cross-modality prediction inconsistency**, defined by the mutual information between predictions from different modalities.

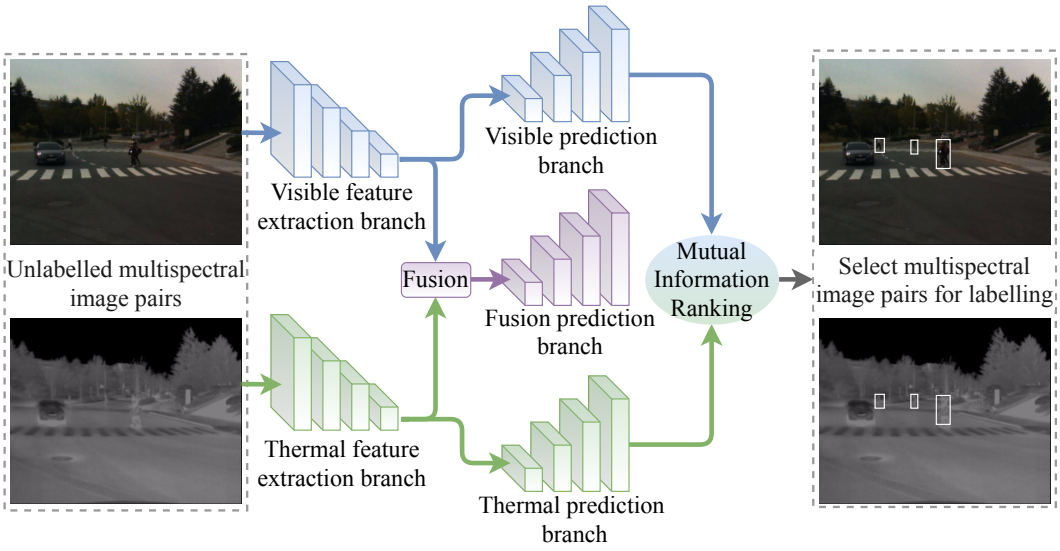


FIGURE 5.2 – Overview of the proposed model for deep active multispectral scene analysis. The blue and green mono-modal branches are used for data informativeness ranking, while the purple one provides the final detection results.

5.1.1 Architecture overview

An overview of our network architecture is given in Figure 5.2. It takes a spatially-aligned multispectral image pair as input, then visible and thermal features are extracted independently via the modality-specific feature extraction networks. Afterwards, three prediction branches are attached: one based on visible features, one based on thermal features, and the last one based on fused features. These three prediction branches are jointly optimized during the model training phase. Here the prediction networks are used for a pedestrian detection task, but can be adapted to other vision tasks such as general object detection or semantic segmentation.

5.1.2 Cross-modality prediction inconsistency

At the selection stage of each active learning cycle, we measure the relevance of labelling a particular image pair by ranking the aforementioned **cross-modality prediction inconsistency**, i.e., we compare predictions from visible and thermal cameras, then select for labelling the image pairs with the highest prediction difference. To be specifically, for each prediction p, its inconsistency is defined as:

$$\mathcal{I} = \mathcal{H}\left(\overline{p}\right) - \frac{1}{2}\sum_{m\in(v,t)}\mathcal{H}\left(p_m\right) \tag{5.1}$$

where p_v and p_t denote the prediction from visible and thermal detection branches; $\overline{p}$ is the average of both predictions; $\mathcal{H}$ is the 2-set entropy function calculated as:

$$\mathcal{H}\left(p\right) = -p\log p - \left(1-p\right)\log\left(1-p\right) \tag{5.2}$$

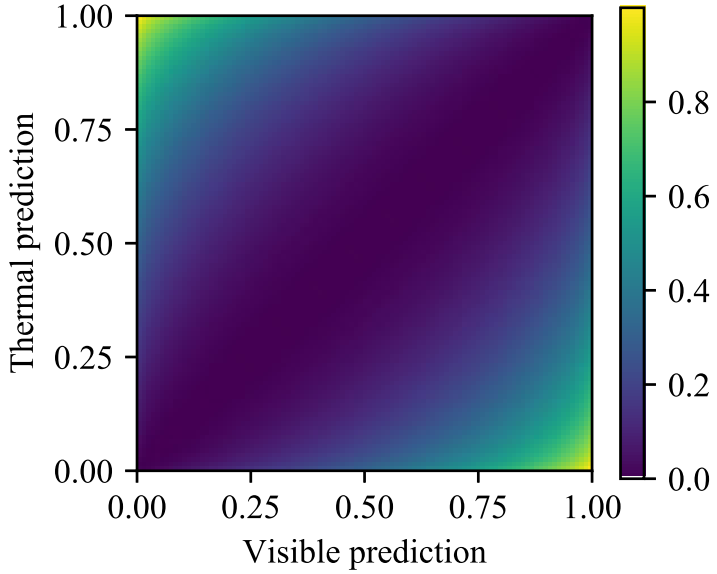


FIGURE 5.3 – Visualization of the proposed cross-modality prediction inconsistency.

For a better understanding of this inconsistency calculation, we plot in Figure 5.3 the visualization of the inconsistency score with different visible (x-axis) and thermal (y-axis) prediction scores. It can be observed that this inconsistency score varies from 0 (very consistent) to 1 (very different).

Scale-balanced inconsistency aggregation. After obtaining the inconsistency for one prediction (i.e. classification of an anchor box for object detection task or classification of a pixel for semantic segmentation task), we adopt the scale-balanced strategy for full-images inconsistency aggregation. This is justified because recent deep learning approaches apply feature pyramid for multiscale prediction. Therefore, if we directly average all predictions for a given image pair, the inconsistency estimation will be dominated by the scale with the most predictions (i.e., the largest feature map in a feature pyramid). Therefore, we first separately average the inconsistency for each pyramid scale, then average the averaged inconsistency across all scales. It could be formulated as:

$$\mathcal{I}_i = \frac{1}{S}\sum_s \frac{1}{P}\sum_p \mathcal{I}_p \tag{5.3}$$

5.1.3 Experimental results

Network architecture. We adopt VGG-16 (Simonyan & Zisserman, 2015) as the feature extraction network, GAFF (as explained in Section 4.4) as the multispectral feature fusion network and SSD (W. Liu et al., 2016) as the prediction network for the object detection tasks. For the semantic segmentation task, the prediction branch is simply one layer of convolution whose number of output channels is equal to the number of classes. In order not to change the aspect ratio of the original images,

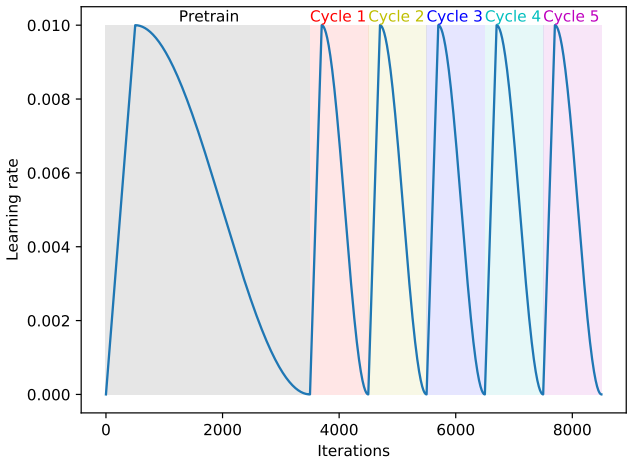


FIGURE 5.4 – Training schedule during the active learning experiment.

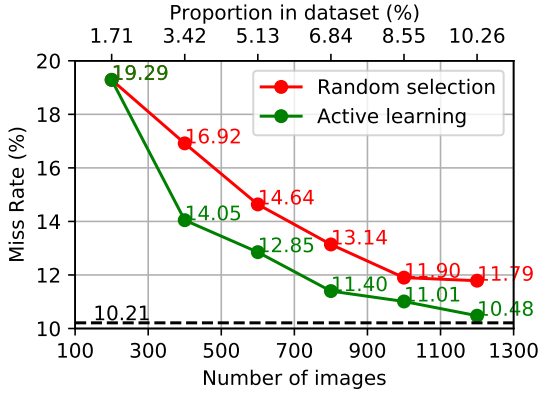
input images are resized to 480×384 or 640×512 for KAIST and FLIR datasets (object detection) and 640×480 for MFNet dataset (semantic segmentation). Random cropping, expanding, flipping are adopted for data augmentation.

Active learning setting. For each active learning experiment, we first randomly initialize a labelled dataset D_l with b images and pretrain the model on D_l; then we actively select b images from an unlabelled dataset D_u with the most cross-modality prediction inconsistency $\mathcal{I}$ for annotation and add these newly labelled images into D_l; afterwards we fine-tune the model with the new D_l; we repeat the previous two steps until the annotation budget B is exhausted. Since semantic segmentation annotations are more difficult to acquire, we set b to 200 and B to 1200 for the object detection tasks, b to 50 and B to 350 for the semantic segmentation task. The training schedule (i.e., learning rate variation) during the whole active learning experiment is plotted in Figure 5.4.

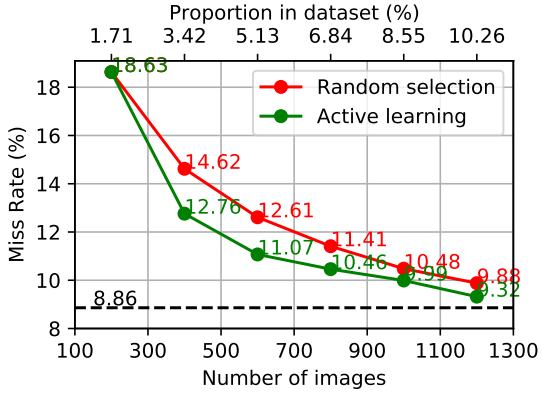
Active vs Random. Figure 5.5 plots the performance evolutions along all learning cycles for KAIST dataset (subfigure a and b), FLIR dataset (c and d) and MFNet dataset (e and f). For all multispectral datasets, all tasks, all evaluation metrics and all input resolutions, our active strategy (green lines in the figure) achieves statistically significant better performance than the random strategy (red lines).

Active vs SotA. We list in Tables 5.1, 5.2 and 5.3 the comparisons between our active learning results and other State-of-the-Art methods for each multispectral dataset. With a small quantity of labelled data (between 10% and 30%), our active models achieve comparable results with fully supervised SotA methods, which demonstrates the effectiveness of our method.

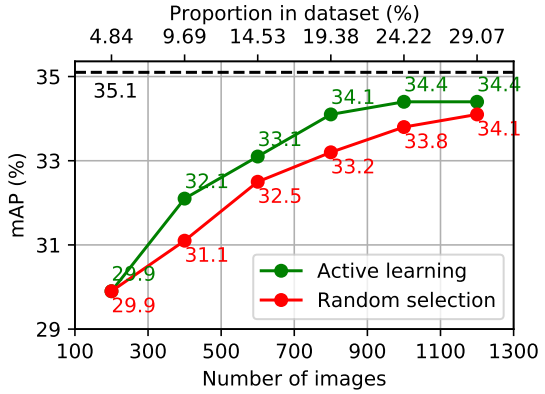
Visualization results. Figure 5.6 shows some image pairs selected by our method. For each dataset, we plot the separate predictions from the visible or thermal cameras, and their cross-modality inconsistency map: our strategy does select some



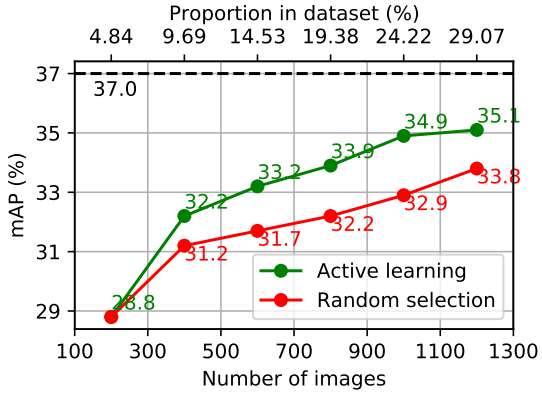
(a) KAIST dataset 480x384



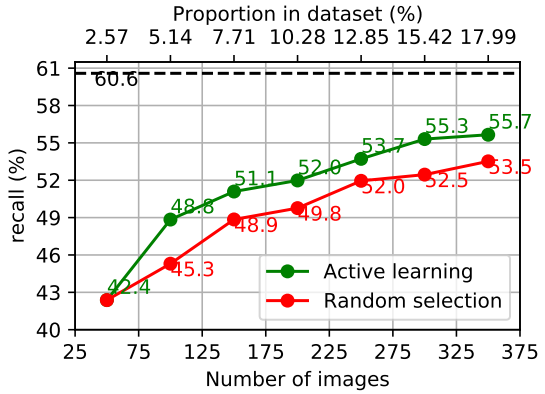
(b) KAIST dataset 640x512



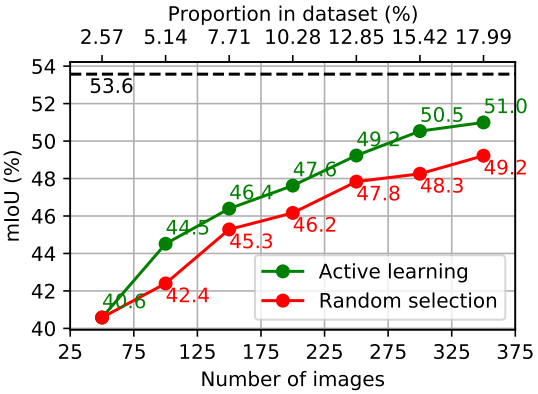
(c) FLIR dataset 480x384



(d) FLIR dataset 640x512



(e) MFNet dataset recall



(f) MFNet dataset mIoU

FIGURE 5.5 – Experimental results of models trained by the proposed active learning strategy (green lines) and random selection strategy (red lines) on KAIST dataset (a, b), FLIR dataset (c, d) and MFNet dataset (e, f). Black dotted lines indicate results trained from full datasets. We conduct experiments with different input image resolutions.

Methods	Miss Rate (lower, better)		
	All	Day	Night
ACF (Hwang et al., 2015)	47.32%	42.57%	56.17%
Halfway Fusion (J. Liu et al., 2016)	25.75%	24.88%	26.59%
Fusion RPN+BF (Konig et al., 2017)	18.29%	19.57%	16.27%
IAF R-CNN (C. Li et al., 2019)	15.73%	14.55%	18.26%
IATDNN+IASS (Guan et al., 2019)	14.95%	14.67%	15.72%
CIAN (L. Zhang, Liu, Zhang, et al., 2019)	14.12%	14.77%	11.13%
MSDS-RCNN (C. Li et al., 2018)	11.34%	10.53%	12.94%
AR-CNN (L. Zhang, Zhu, et al., 2019)	9.34%	9.94%	8.38%
MBNet (K. Zhou et al., 2020)	8.13%	8.28%	7.86%
Ours (full dataset)	8.86%	10.01%	6.77%
Ours (10.26% of data)	9.32%	10.13%	7.70%

TABLE 5.1 – Comparison between state-of-the-art multispectral pedestrian segmentation methods and ours on KAIST dataset (Hwang et al., 2015).

Methods	mAP	AP50	AP75
GAFF (H. Zhang et al., 2021a)	37.3%	72.7%	30.9%
Ours (full dataset)	37.0%	72.1%	31.2%
Ours (29.07% of data)	35.1%	71.0%	30.6%

TABLE 5.2 – Comparison between state-of-the-art multispectral object detection methods and ours on FLIR dataset.

Methods	mIoU (higher, better)		
	All	Day	Night
MFNet (Ha et al., 2017)	39.7%	36.1%	36.8%
FuseNet (Hazirbas et al., 2016)	45.6%	41.0%	43.9%
RTFNet-50 (Y. Sun et al., 2019)	51.7%	44.4%	52.0%
RTFNet-152 (Y. Sun et al., 2019)	53.2%	45.8%	54.8%
Ours (full dataset)	53.6%	46.8%	53.3%
Ours (17.99% of data)	51.0%	46.6%	48.9%

TABLE 5.3 – Comparison between state-of-the-art multispectral semantic segmentation methods and ours on MFNet dataset (Ha et al., 2017).

difficult cases where at least one modality makes mistakes. **We believe that adding these informative examples into the labelled dataset for fine-tuning is the main reason for performance improvements.**

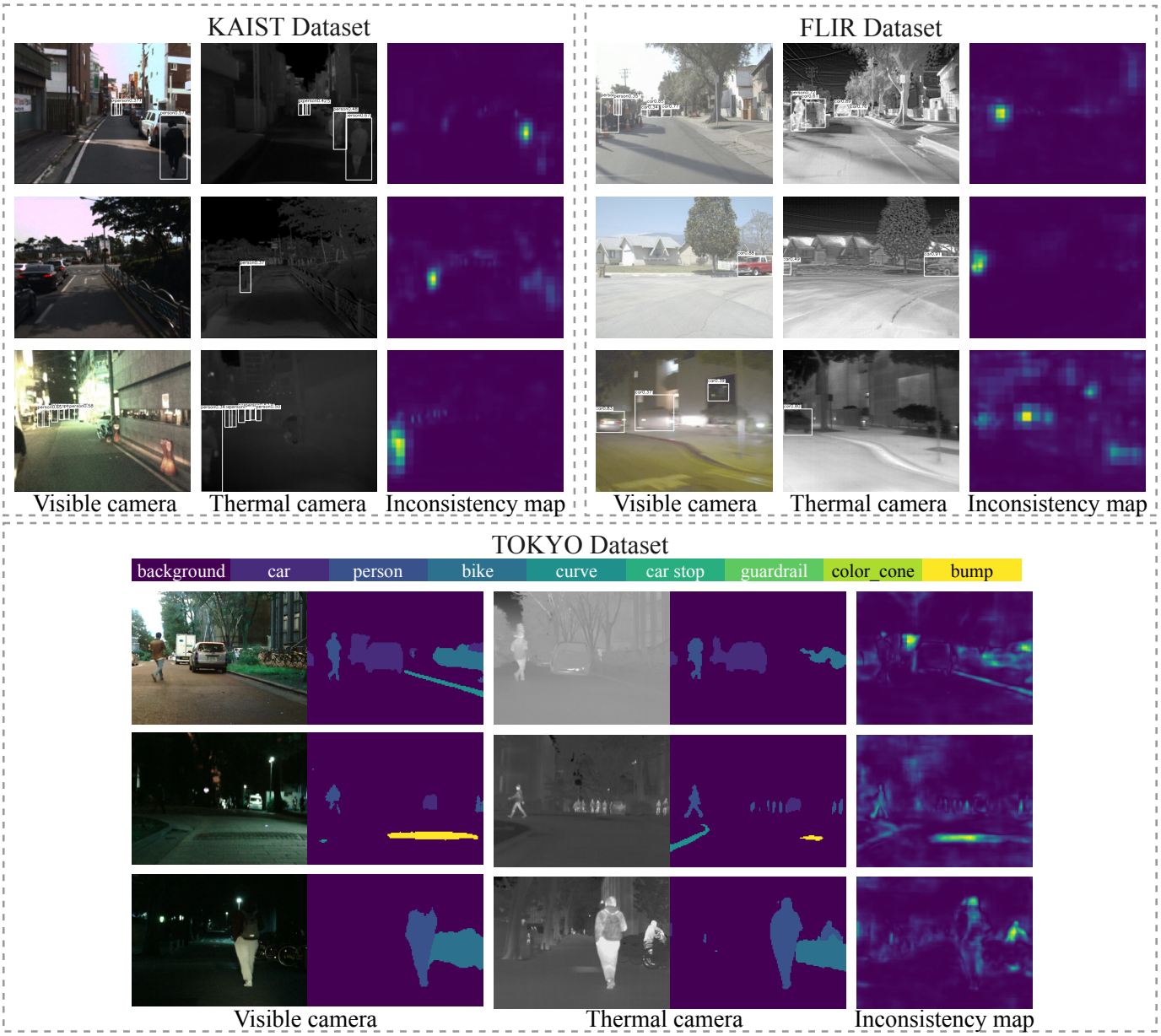


FIGURE 5.6 – Examples of selected image pairs for labelling by the proposed method. Zoom in to see details.

5.2 Low-cost Multispectral Scene Analysis with Modality Distillation

Under the conventional settings of multispectral scene analysis, thermal cameras and visible ones must provide image pairs with identical perception fields and identical spatial resolution. The former requirement can be achieved through camera calibration. However, due to the extreme price gap between high-resolution visible and thermal cameras[1], the requirement of identical spatial resolution usually leads to either 1) RGB image downsampling that may cause information loss or 2) high

[1]A typical thermal camera of resolution 640 $\times$ 480 could cost more than 8,000 USD. When the resolution is reduced to 80 $\times$ 60, the price becomes much more affordable (around 200 USD).

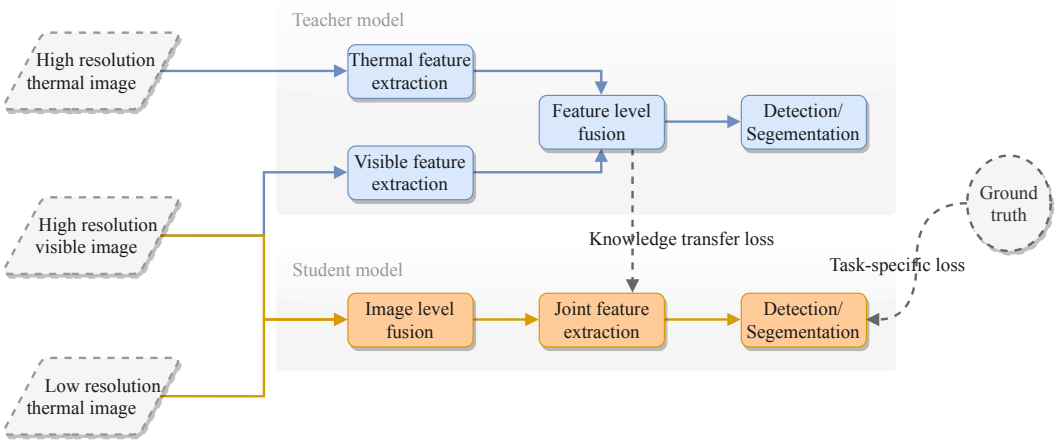


FIGURE 5.7 – Overview of the proposed **Modality Distillation** (MD) framework. Blue and yellow blocks represent components from teacher and student models.

manufacturing costs for thermal cameras that prevent massive production. From a practical point of view, using a high-resolution visible camera and a low-resolution thermal one would be the best compromise in performance/price.

Another constraint from the current multispectral systems lies in the software part. Nowadays, deep learning-based methods dominate the field of multispectral scene analysis. As explained in Section 2.2, multispectral information fusion methods can be categorized into: image-level fusion, feature-level fusion or decision-level fusion. Architectures that implement a feature-level fusion, usually adopt a two-stream neural network (one network to each source), have been proven to outperform the other strategies, and are currently the most studied in the literature (Ha et al., 2017; Hazirbas et al., 2016; Konig et al., 2017; C. Li et al., 2018; J. Liu et al., 2016; Y. Sun et al., 2019; L. Zhang, Liu, Zhang, et al., 2019; K. Zhou et al., 2020). However, since two-stream networks duplicate the number of parameters and calculations of the backbone subnetwork, the computational overhead is huge compared to one-stream network, which is particularly undesirable for software deployment on embedded devices.

To tackle the aforementioned hardware and software constraints, we propose a novel knowledge distillation framework named **Modality Distillation** (MD). This framework follows two steps: Firstly, a multispectral system with high-resolution visible and thermal cameras is used to collect training data and to learn a precise but complex two-stream neural network for scene analysis. This model will be used as a teacher model with fixed weights. Secondly, a more efficient image-level fusion student model is trained with high-resolution RGB images and downsampled thermal images to simulate production systems that are equipped with more economical low-resolution thermal cameras. The knowledge learnt from the teacher model is transferred to the student model to mimic the more accurate feature-level fusion architecture and to reconstruct high-resolution thermal details.

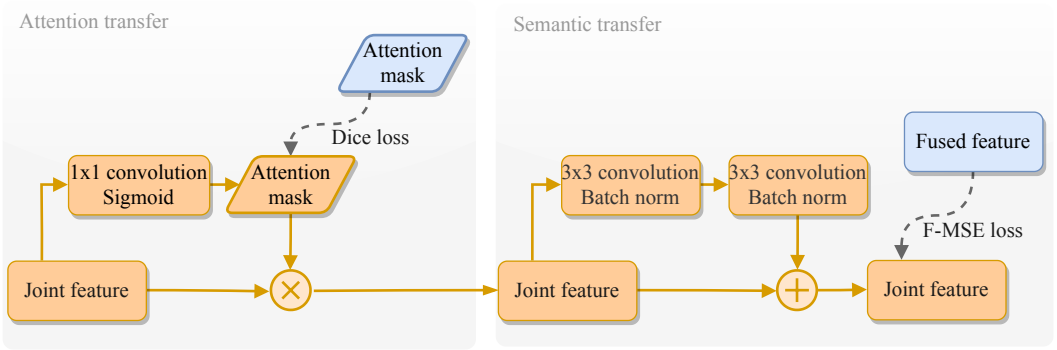


FIGURE 5.8 – Details on knowledge transfer modules. Blue and yellow blocks represent components from teacher and student models.

5.2.1 Architecture overview

As illustrated in Figure 5.7, the proposed **Modality Distillation** (MD) framework includes a teacher model (upper model in blue) and a student model (lower model in yellow). The teacher model takes **high-resolution** multispectral image pairs as input, and employs a **two-stream architecture** consisting of: two separate feature extraction networks, a GAFF module (explained in Section 4.4) for multispectral fusion and a task-specific network for pedestrian detection/semantic segmentation. Contrarily to the teacher model, the student model uses a **low-resolution** thermal input and a **one-stream** feature extraction network that takes as input the image level fusion of both modalities (through different input channels). We also conduct distillation experiments for a student model without the thermal modality, i.e. in this particular case, we attempt to use a multispectral teacher to improve the performance of a visible-only student.

The proposed MD framework includes two training stages. In the first stage, we train the teacher model and fix its weights, such that the fused features from GAFF module contain the rich semantics of high-resolution thermal-visible image pairs. These features are used to guide the training of the student model; In the second stage, the optimization of the student model is supervised by a task-specific loss (e.g., pedestrian detection or semantic segmentation loss) as well as the knowledge transfer loss. The objective of the knowledge transfer loss is two-fold: using a more efficient one-stream network to mimic a more precise two-stream network and using the more available low-resolution thermal images to reconstruct high-resolution thermal details. Finally, we obtain a student model that takes low-resolution thermal images as input, and the required parameters and calculations are greatly reduced. Meanwhile, the precision of the low thermal resolution one-stream student model is supposed to be close to the high thermal resolution two-stream teacher.

5.2.2 Knowledge transfer modules

To preserve the knowledge learnt from the teacher model to the maximum extent, we apply two knowledge transfer strategies: **Attention transfer** that guides the one-stream model to mimic the two-stream attentive fusion, and **Semantic transfer** that rebuilds high-resolution visual details from low-resolution thermal images.

Attention transfer. In the two-stream teacher model, GAFF significantly improves the scene analysis performance. However, such a multispectral feature fusion module does not exist in a one-stream student model. Thus, as illustrated in the left part of Figure 5.8, we design the **Attention transfer** module to simulate this attentive fusion in a one-stream model. The teacher attention mask is the combination of intra- and inter-modality attention masks from visible and thermal modalities[2]. To keep the architecture simple, the student attention mask is generated by a 1×1 convolution followed by a Sigmoid activation, and is supervised by minimizing the Dice loss (Dice, 1945) between the student and teacher attention masks:

$$\begin{aligned} m_{teacher} &= m_{intra}^{visible} \otimes m_{inter}^{visible} + m_{intra}^{thermal} \otimes m_{inter}^{thermal} \\ m_{student} &= \mathcal{F}(f_{student}) \\ L_{attention} &= 1 - \frac{2|m_{student} \otimes m_{teacher}|}{|m_{student}| + |m_{teacher}|} \end{aligned} \tag{5.4}$$

where $L_{attention}$ denotes the attention transfer loss; $m_{teacher}$ and $m_{student}$ represent the teacher and student attention masks respectively; $f_{student}$ denotes the student features acquired from the joint feature branch; $\mathcal{F}$ represents a 1×1 convolution followed by a Sigmoid activation; $\otimes$ and $||$ represent respectively the pixel-wise multiplication and summation operation.

Semantic transfer. To compensate for the resolution reduction of thermal input images, the **Semantic transfer** module performs an implicit super-resolution of student feature maps. As shown in the right part of Figure 5.8, we use a basic residual block (He et al., 2016) to increase the details in the joint features. **Semantic transfer** aims to minimize the distance between the student (joint) and teacher (fused) feature maps. However, optimizing this distance has proven to be difficult. At first glance, this is due to the extreme imbalance between the foreground and background areas. Inspired by the Focal loss (Lin, Goyal, et al., 2017), we argue that the true problem lies in the extreme imbalance between easily-mimic and hardly-mimic areas. Therefore, we propose the Focal Mean Square Error (F-MSE) loss defined as:

$$\begin{aligned} d &= (f_{student} - f_{teacher})^2 \\ L_{semantic} &= \sum_w \sum_h \frac{1}{n} (\delta(\sum_n d) \times \sum_n d) \end{aligned} \tag{5.5}$$

where $L_{semantic}$ denotes the semantic transfer loss; d is the squared L2 distance between student and teacher feature maps. δ signifies the Softmax function; w, h, n represent the width, height and depth of feature maps, respectively.

The only difference between the proposed F-MSE loss and the standard MSE loss is the spatial re-weighting based on the feature-mimicking error. Concretely, the Softmax function generates a 2-D re-weighting mask, where each value reflects

[2]We refer the reader to Section 4.4 for more details about GAFF module.

the difficulty of feature mimicking on a specific area, and the sum of all values on the mask is equal to 1. In such a manner, the optimization adaptively "focuses" on mis-predicted areas, and the imbalance problem is therefore solved.

5.2.3 Experimental results

Network architecture. For all experiments, we apply ResNet-18 (He et al., 2016) as the feature extraction network, RetinaNet (Lin, Goyal, et al., 2017) as the pedestrian detection network and PSPNet (Zhao et al., 2017) as the semantic segmentation network. For the teacher model (Figure 5.7 upper model), GAFF module is adopted for the attentive fusion of visible and thermal features. For the student model (Figure 5.7 lower model), visible and thermal images are concatenated to generate 6-channel multispectral inputs, i.e., 3 channels from each modality. The first convolution layer is modified to suit the 6-channel input[3]. Note that instead of generating 4-channel input as done in (Wagner et al., 2016), we duplicate the single-channel thermal images into 3 channels to balance the contribution of visible and thermal spectrum in the first convolution layer.

Input resolution. The resolution of visible input images are set identical to previous methods for fair comparisons. More specifically, the resolution is 640×512 on KAIST dataset and 640×480 on MFNet dataset. To simulate the low-resolution thermal camera in actual products, we downsample the high-resolution thermal images through bilinear interpolation. These downsampled small thermal images are then re-scaled to the original spatial size to concatenate with the RGB images. Note that the high-resolution thermal details are already lost in the first interpolation operation. Considering the camera price and the total number of pixels, 16 times thermal resolution downsampling is regarded as the most **practical** case (e.g., downsampling from 640×512 to 160×128).

Baseline results. Image-level, feature-level and decision-level are the three major fusion methods for multispectral scene analysis. We list in Table 5.4 and 5.5 their prediction accuracy and inference time[4] on KAIST dataset (Hwang et al., 2015) and on MFNet dataset (Ha et al., 2017), respectively. The visible-only results are also listed for reference. For simplicity, we average the prediction from visible and thermal images for decision-level fusion. The tables show that, regardless of the information fusion stage (image-level, feature-level or decision-level), multispectral methods greatly improve the detection/segmentation accuracy compared to the visible-only model, especially for nighttime detection/segmentation. Feature-level and decision-level fusion methods almost double the execution runtimes as well as the number of parameters of a visible-only model. In contrast, the computational overhead for the image-level fusion is negligible, which shows the relevance of this fusion method when fewer computational resources are available.

[3]Concretely, the pretrained ImageNet (Deng et al., 2009) weights for the first convolution layer are duplicated along the input channel dimension, and the values are halved. The bias values and the following batch normalization parameters remain unchanged.

[4]The runtimes are measured on an Nvidia GTX 1080Ti GPU.

Fusion stage	Miss Rate			Parameters	Runtime
	Day	Night	All		
Visible-only	16.95%±0.88	35.15%±1.20	22.84%±0.77	**1.244e+7**	**6.48ms**
Image-level	10.73%±0.44	6.61%±0.28	9.40%±0.39	1.245e+7	6.55ms
Feature-level	**9.37%±0.17**	**4.71%±0.34**	**7.77%±0.07**	2.377e+7	10.97ms
Decision-level	10.74%±0.21	9.25%±0.62	10.34%±0.32	2.488e+7	12.96ms

TABLE 5.4 – Different fusion methods on KAIST dataset. For fair comparisons, all listed methods use the same feature extraction network (ResNet-18) and detection network (RetinaNet).

Fusion stage	Mean Accuracy			Parameters	Runtime
	Day	Night	All		
Visible-only	50.83%±0.21	52.26%±0.53	55.06%±0.21	**1.138e+7**	**4.57ms**
Image-level	54.97%±0.50	56.07%±0.21	59.42%±0.22	1.139e+7	4.68ms
Feature-level	**57.21%±0.23**	**62.18%±0.64**	**63.45%±0.24**	2.270e+7	8.94ms
Decision-level	51.72%±0.19	53.37%±0.30	56.21%±0.14	2.276e+7	9.14ms

TABLE 5.5 – Different fusion methods on MFNet dataset. For fair comparisons, all listed methods use the same feature network (ResNet-18) and segmentation network (PSPNet).

Thermal resolution	MD	Miss Rate		
		Day	Night	All
640 × 512 (full resolution)		10.73%±0.44	6.61%±0.28	9.40%±0.39
	✓	**9.45%±0.49**	**4.61%±0.20**	**7.78%±0.28**
320 × 256 (4x downsample)		11.57%±0.75	6.51%±0.54	9.84%±0.72
	✓	**9.39%±0.27**	**5.07%±0.43**	**7.91%±0.11**
160 × 128 (16x downsample)		12.09%±0.24	6.73%±0.55	10.17%±0.42
	✓	**9.85%±0.21**	**4.84%±0.26**	**8.03%±0.19**
80 × 64 (64x downsample)		14.92%±0.47	10.66%±0.62	13.37%±0.30
	✓	**10.75%±0.10**	**7.07%±0.21**	**9.50%±0.06**
Visible-only		16.95%±0.88	35.15%±1.20	22.84%±0.77
	✓	**14.74%±0.45**	**34.13%±0.49**	**21.08%±0.21**

TABLE 5.6 – Comparison between native models and distilled models on KAIST dataset under different thermal resolution settings (from full thermal resolution to no thermal scenario). All listed methods use an image-level fusion architecture.

Distillation results. We list in Table 5.6 and 5.7 the comparisons between native image-level fusion models and distilled image-level fusion models (i.e., student models) on KAIST dataset (Hwang et al., 2015) and MFNet dataset (Ha et al., 2017), respectively. It can be observed that MD strategy brings important improvements for all thermal resolutions for both datasets.

Thermal resolution	MD	Mean Accuracy		
		Day	Night	All
640 × 480 (full resolution)		54.97%±0.50	56.07%±0.21	59.42%±0.22
	✓	**59.71%±0.31**	**62.78%±0.28**	**64.93%±0.11**
320 × 240 (4x downsample)		53.89%±0.69	55.18%±0.11	57.93%±0.27
	✓	**58.32%±0.58**	**61.88%±0.52**	**64.25%±0.11**
160 × 120 (16x downsample)		53.85%±0.62	55.43%±0.45	58.21%±0.46
	✓	**58.46%±0.15**	**61.37%±1.09**	**63.52%±0.87**
80 × 60 (64x downsample)		53.68%±0.07	53.68%±0.33	57.06%±0.18
	✓	**57.58%±0.44**	**59.67%±0.82**	**62.62%±0.81**
Visible-only		50.83%±0.21	52.26%±0.53	55.06%±0.21
	✓	**57.74%±0.13**	**56.67%±0.64**	**60.62%±0.42**

TABLE 5.7 – Comparison between native models and distilled models on MFNet dataset under different thermal resolution settings (from full thermal resolution to no thermal scenario). All listed methods use an image-level fusion architecture.

Specifically, on the multispectral pedestrian detection task (Table 5.6), our full thermal resolution result with MD is already close to that of the feature-level fusion model (i.e., teacher model) shown in Table 5.4 (7.78% versus 7.77%), while the inference time is almost halved (10.97ms versus 6.55ms). When it comes to the most practical case where thermal resolution is 16 times lower than visible resolution, MD strategy brings 2.14% of Miss Rate improvement, and the performance difference compared to the teacher model is only 0.26% (8.03% versus 7.77%). We show some detection results from native model and distilled model for this practical case in Figure 5.9, and it can be observed that our distilled model provides more precise detection results. Interestingly, the nighttime detection precision is boosted by 27.06% (7.07% versus 34.13%) even if the thermal resolution is reduced to 80 × 64 (i.e., 64 times downsampled), proving the necessity of the thermal modality in nighttime detections. Moreover, our strategy remains helpful when the thermal image is completely removed (e.g., the Miss Rate for visible-only model is reduced from 22.84% to 21.08%). Here, the multispectral knowledge from the teacher model allows the visible-only student to perform pseudo-multispectral detection, which is the main reason of improvements.

On the multispectral semantic segmentation task, the improvements are more important (around 5% for all thermal resolutions using MD). It is noteworthy that the performance of the distilled visible-only model is even better than that of the native full-resolution image-level fusion model (60.62% versus 59.42%). Here, our assumption is that the multispectral semantic segmentation task is more critical for the choice of fusion architecture, e.g., according to Tab. 5.5, native image-level fusion performs 4.03% worse than feature-level fusion. This may be the reason why rare previous work use image-level fusion for multispectral semantic segmentation. However, our MD strategy makes the student model mimic a feature-level

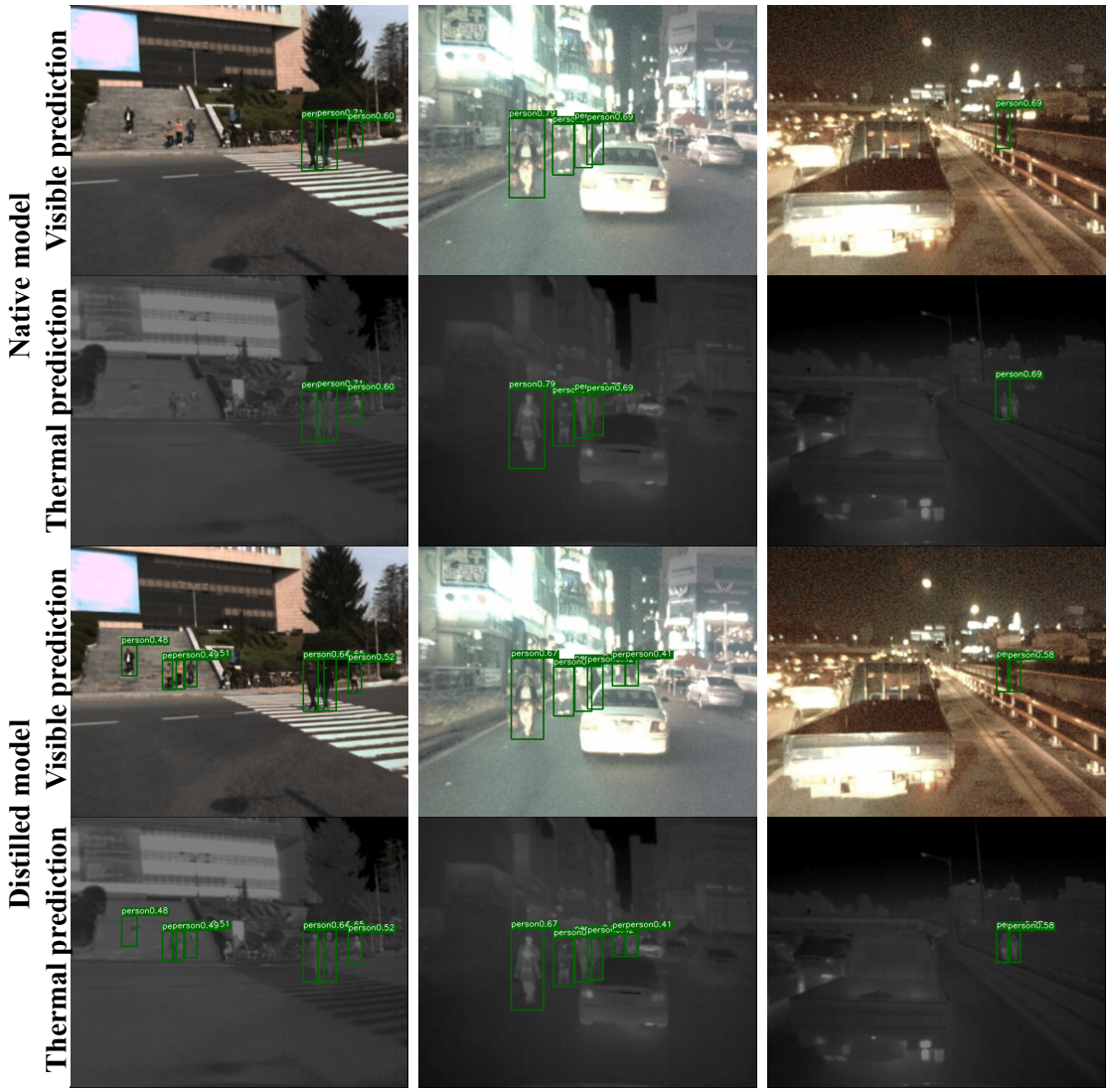


FIGURE 5.9 – Visual improvements on KAIST dataset.

fusion teacher model, which compensates its established disadvantage of image-level fusion architecture, and thus brings tremendous accuracy improvements. For the practical case (16 times thermal resolution downsampling), the mean accuracy difference with the teacher model is minor (63.52% versus 63.45%). We visualize in Figure 5.10 the segmentation results from the native model and our distilled model for the practical case, and it could be noted that the improvement from MD strategy on segmentation quality is obvious.

Comparing with state-of-the-art. We compare the results of our distilled models (which adopt image-level fusion) with state-of-the-art methods (all adopting feature-level fusion) on KAIST dataset (Table 5.8) and MFNet dataset (Table 5.9). Note that our teacher models use GAFF fusion module, and this method has already been shown to give better results than its competitors. However, our goal here is to show how our student models (which are supposed to be less good than their teachers) perform compared to their competitors. Specifically, we provide our

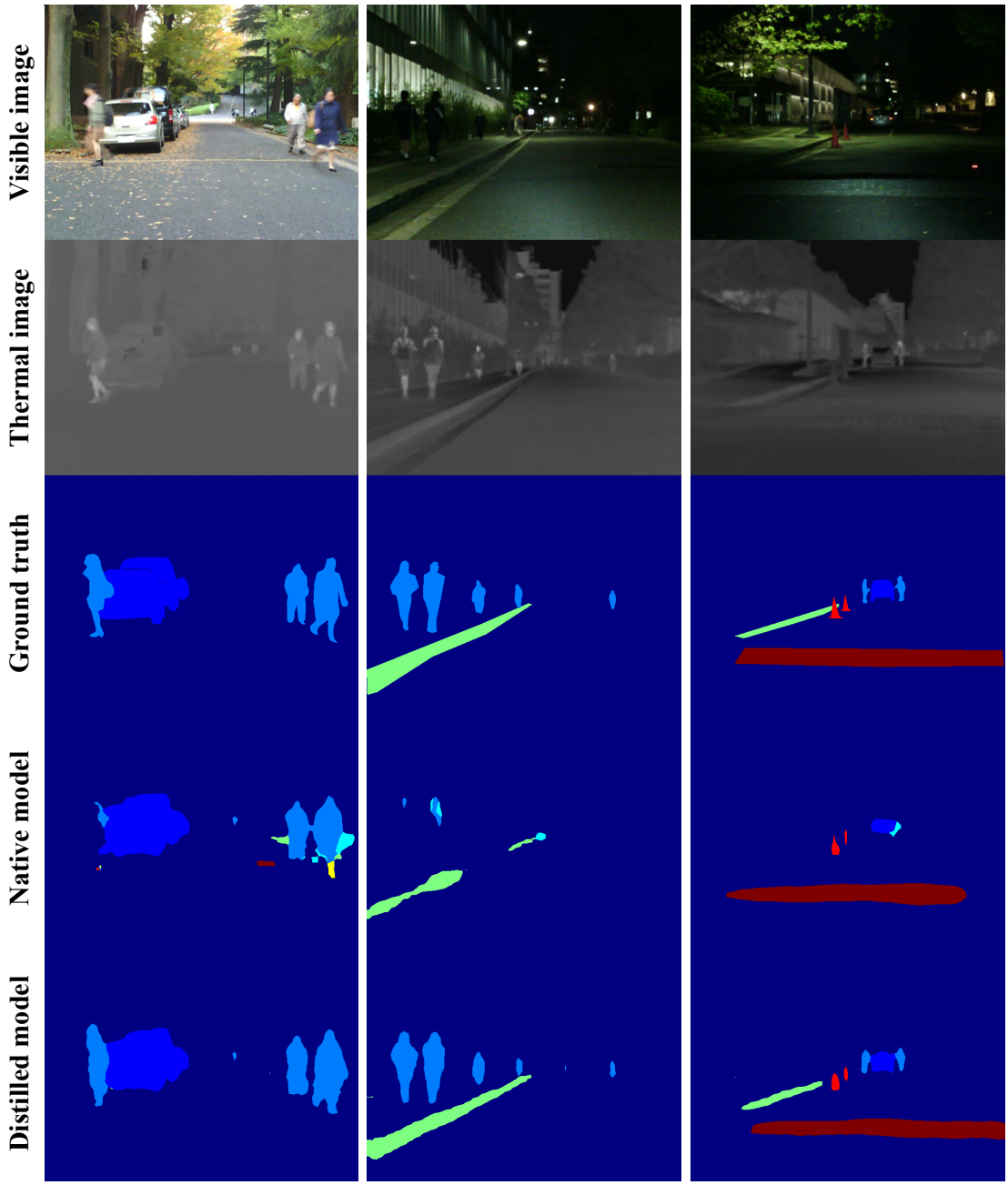


FIGURE 5.10 – Visual improvements on MFNet dataset.

one-stream student models' results using full thermal resolution (same condition as our competitors, denoted as "full") and using 16 times downsampled thermal resolution (denoted as "practical"). We also list our two-stream teacher models' results (denoted as "teacher") for reference. We consider "practical" the most interesting setting for actual multispectral applications.

On the multispectral pedestrian detection task (Table 5.8), the achieved Miss Rate from the distilled "practical" model is already better than that of the best feature-level fusion methods in the literature (K. Zhou et al., 2020) (8.03% versus 8.13%). It should be noted that our "practical" model takes downsampled thermal images as input and adopts a much simpler architecture (one-stream networks for

Method	Miss Rate			Runtime
	Day	Night	All	
ACF (Hwang et al., 2015)	42.57%	56.17%	47.32%	2730ms
Halfway Fusion (J. Liu et al., 2016)	24.88%	26.59%	25.75%	430ms
FusionRPN+BF (Konig et al., 2017)	19.57%	16.27%	18.29%	800ms
IAF R-CNN (C. Li et al., 2019)	14.55%	18.26%	15.73%	210ms
IATDNN+IASS (Guan et al., 2019)	14.67%	15.72%	14.95%	250ms
RFA (L. Zhang, Liu, Chen, et al., 2019)	16.78%	10.21%	14.61%	80ms
CIAN (L. Zhang, Liu, Zhang, et al., 2019)	14.77%	11.13%	14.12%	70ms
MSDS-RCNN (C. Li et al., 2018)	10.53%	12.94%	11.34%	220ms
AR-CNN (L. Zhang, Zhu, et al., 2019)	9.94%	8.38%	9.34%	120ms
MBNet (K. Zhou et al., 2020)	**8.28%**	7.86%	8.13%	70ms
Ours (teacher)	9.37%	4.71%	**7.77%**	11ms
Ours (full)	9.45%	**4.61%**	7.78%	**7ms**
Ours (practical)	9.85%	4.84%	8.03%	**7ms**

TABLE 5.8 – Comparison between state-of-the-art multispectral pedestrian detection methods and ours on KAIST dataset. Our competitors' results are taken from (K. Zhou et al., 2020).

"full"/"practical" and two-stream networks for others). It is also worth noting that our nighttime detection performance surpasses all previous methods, which proves that the thermal information has been well-preserved in the student model.

On the multispectral semantic segmentation task (Table 5.9), thanks to the substantial accuracy improvements from MD (about 5%), our distilled "practical" model also surpasses the best previous result (Y. Sun et al., 2019) (63.5% versus 63.1%). For this dataset as well, all our trained models (including the "practical" model with downsampled thermal input) have obvious advantage in nighttime prediction. It should be pointed out that both our distilled models with one-stream ResNet-18 feature network even outperform RTFNet-152 (Y. Sun et al., 2019) with two-stream ResNet-152 feature network, demonstrating the high efficiency of our distilled models (ours are about 6 times faster than RTFNet-152). More surprisingly, we can see in Table 5.9 that the full student model gives slightly better results than the teacher model. The student model's feature extraction network is the same as the teacher's one, so the student could theoretically achieve similar performance, and the student model has more sources of supervision, i.e., the ground truth and the knowledge learnt from the teacher model, which we believe is the reason for the higher performance shown by the student model.

Ablation experiments. To explore the effects of the proposed **Attention transfer** and **Semantic transfer** modules, we conduct ablation experiments on KAIST dataset (Table 5.10) and MFNet dataset (Table 5.11), under the most practical case (thermal images are 16 times downsampled). The "S" and "A" denote **Semantic transfer** and **Attention transfer** modules as illustrated in Figure 5.8 right and left parts, respectively. We conduct comparative experiments between the traditional MSE loss

Method	Mean Accuracy			Runtime
	Day	Night	All	
MFNet (Ha et al., 2017)	42.6%	41.4%	45.1%	4.35ms
FuseNet (Hazirbas et al., 2016)	49.5%	48.9%	52.4%	**3.92ms**
RTFNet-50 (Y. Sun et al., 2019)	57.3%	59.4%	62.2%	11.25ms
RTFNet-152 (Y. Sun et al., 2019)	**60.0%**	60.7%	63.1%	29.35ms
Ours (teacher)	57.2%	62.2%	63.5%	8.94ms
Ours (full)	59.7%	**62.8%**	**64.9%**	4.92ms
Ours (practical)	57.6%	61.4%	63.5%	4.92ms

TABLE 5.9 – Comparison between state-of-the-art multispectral semantic segmentation methods and ours on MFNet dataset. Our competitors' results are taken from (Y. Sun et al., 2019).

S(M)	S(F)	A	Miss Rate		
			Day	Night	All
			12.09%±0.24	6.73%±0.55	10.17%±0.42
✓			11.92%±0.21	6.69%±0.33	10.09%±0.06
	✓		10.24%±0.30	6.93%±0.08	9.11%±0.15
		✓	10.61%±0.43	5.83%±0.39	8.99%±0.23
	✓	✓	**9.85%±0.21**	**4.84%±0.26**	**8.03%±0.19**

TABLE 5.10 – Ablation experiments on KAIST dataset. We study the effects of Semantic transfer (with MSE or F-MSE loss) and Attention transfer modules in the proposed MD framework.

S(M)	S(F)	A	Mean Accuracy		
			Day	Night	All
			53.85%±0.62	55.43%±0.45	58.21%±0.46
✓			56.04%±0.66	57.87%±0.23	60.41%±0.28
	✓		57.49%±0.47	58.47%±0.41	61.16%±0.13
		✓	57.55%±0.65	58.46%±0.29	61.51%±0.30
	✓	✓	**57.58%±0.44**	**61.37%±1.09**	**63.52%±0.87**

TABLE 5.11 – Ablation experiments on MFNet dataset. We study the effects of Semantic transfer (with MSE or F-MSE loss) and Attention transfer modules in the proposed MD framework.

(denoted as "M") and the proposed F-MSE loss (denoted as "F") in the **Semantic transfer** module. According to our experimental results, the latter provides better performance. Moreover, we visualize some examples of the 2-D spatial re-weighting mask from F-MSE loss (Equation 5.5) in Figure 5.11, and it can be observed that the optimization on the teacher-student feature mimicking is automatically "focused" on more important areas, e.g., pedestrians, vehicles and colour cones. This specific loss tackles the imbalance problem in the **Semantic transfer** module. In conclusion,

FIGURE 5.11 – Visualization of the visible-thermal image pairs and the 2-D spatial re-weighting masks from the proposed F-MSE loss. The first two lines of multispectral images pairs come from KAIST dataset, and the last two lines come from MFNet dataset.

according to our ablation experiments on the two datasets, both the proposed **Semantic transfer** (S(F)) and **Attention transfer** (A) modules bring notable improvements and their combination leads to the best performance.

Chapter 6

Conclusions and future works

Contents

6.1 Conclusions

In this thesis, we investigated three main challenges associated with multispectral scene analysis: (1) the fast and accurate detection of objects of interest; (2) the dynamic and adaptive fusion of information from different modalities; (3) low-cost and low-energy multispectral object detection and the reduction of its manual annotation efforts. Specifically, we provided various solutions through three vision-based scene analysis applications, i.e., object detection from RGB images, object detection from multispectral image pairs and semantic segmentation from multispectral image pairs. In this thesis, we first presented our application context and related research background in Chapter 1 and 2. Then, for each of the three challenges, we elaborated one chapter to introduce our proposed methods and achieved results.

To cope with the first challenge, we optimized both the precision and the speed of existing object detection models. To this end, various best practices for model training and two solutions from different perspectives were proposed. We first introduced a novel label assignment strategy named **Mutual Guidance**, which assigns labels for the classification task according to the prediction quality on the localization task and vice versa. This strategy not only provides an adaptive matching between anchors and objects, but also tackles the prediction misalignment problem between localization and classification tasks. Our second contribution is to introduce a novel model compression method named **PDF-Distil**, which leverages the teacher-student prediction disagreements to guide the knowledge transfer in a feature-based detection distillation framework. By incorporating the logits-level information, the feature mimicking is automatically focused on areas where the student model makes inaccurate predictions, thereby greatly reduces the computational complexity of object detection models.

To study the second challenge, we attempted to deal with the modality inconsistency problem when performing multispectral information fusion. Information from visible and thermal cameras are complementary, but the multispectral fusion becomes difficult when the two cameras provide contradictory information. Three solutions were proposed to deal with this situation: First, we proposed a novel fusion network named **Cycle Fuse-and-Refine**, which consecutively refines the monospectral features with the fused multispectral features during the fusion process. This cascaded architecture reduces the difference between visible and thermal features, and improves the overall feature quality. We then propose a second multispectral fusion module named **Progressive Spectral Fuse**, where multispectral features are progressively fused throughout multiple convolution levels. The third contribution based on supervised attention mechanism is subsequently proposed. When applying attention mechanism for multispectral fusion, we expect the network to actively select the modality with superior feature quality. However, we argued that the lack of guidance is a limitation for the attention-based fusion, and we proposed **Guided Attentive Feature Fusion** to explicitly guide this fusion process. Without handcrafted assumptions or additional annotations, our method realizes a fully adaptive fusion of visible and thermal features.

Regarding the third challenge, we intended to integrate active learning and knowledge distillation into the multispectral scene analysis framework. We studied the complementarity between multispectral cameras for the **active selection** of multispectral image pairs to annotate. Different from previous active learning methods where only RGB images are considered, we take the prediction difference between two sensors into account. Moreover, in order to reduce the hardware cost of multispectral scene analysis systems, we propose a novel knowledge distillation framework named **Modality Distillation**, which distils the knowledge from a high thermal resolution two-stream network to a low thermal resolution one-stream network. The distilled model could perform precise prediction on widely available low-resolution thermal cameras, and shows a similar computational complexity to the RGB-only models.

6.2 Application to remote sensing data

To emphasize the possible impact of our methods outside the intelligent surveillance use case, we apply our proposed contributions to a remote sensing scenario.. Remote sensing is an important technology for earth observation. The principle is to analyse images captured from airborne or satellite sensors. These images taken from a high altitude allow us to efficiently detect buildings, roads, vehicles or any other objects of interest for earth observation. Specifically, we focus on the vehicle detection task on the VEDAI dataset (Razakarivony & Jurie, 2016).

VEDAI dataset. VEDAI is the short for VEhicle Detection in Aerial Imagery. This dataset targets the detection of vehicles from eight subdivided categories, including boat, camping car, car, pickup, plane, tractor, truck and vans. It contains 1,125

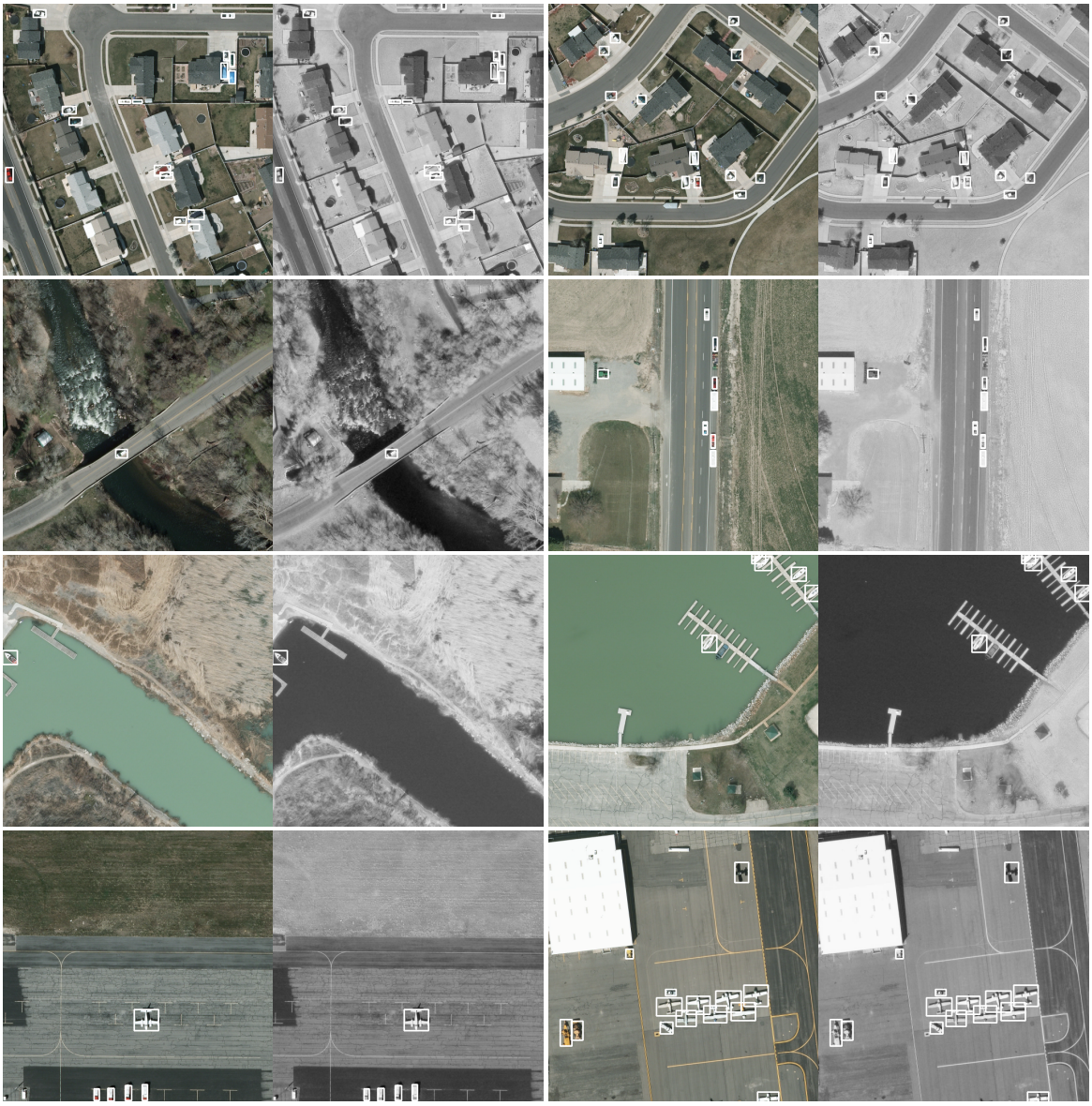


FIGURE 6.1 – Example of multispectral aerial images from VEDAI dataset.

and 121 aerial images for training and evaluation, respectively. Moreover, each provided aerial image is available in both visible and thermal spectral bands. Figure 6.1 shows some examples of multispectral image pairs from this dataset. As is shown, the aerial images are taken at different places, such as villages, roads, docks and airports, and the targeting vehicles exhibit different variabilities, such as multiple orientations, lighting/shadowing changes or occlusions. Following the conventional practices, the mean Average Precision (mAP) is used as the evaluation metric, where the IoU threshold is set as 50% ($AP50$).

Network architectures. Experiments are conducted with ResNet backbone (He et al., 2016) and RetinaNet detector (Lin, Goyal, et al., 2017). According to our observations, vehicles in VEDAI dataset are generally similar in size (as shown in Figure

Model	MG	Fitnets	PDF-Distil	*mAP*
RetinaNet with ResNet-34 backbone (teacher)				77.0
	✓			77.6 (+0.6)
RetinaNet with ResNet-18 backbone (student)				72.8
	✓			75.1 (+2.3)
	✓	✓		75.3 (+2.5)
	✓		✓	**78.8 (+6.0)**

TABLE 6.1 – Experimental results for visible-only vehicle detection.

Model	MG	Fitnets	PDF-Distil	*mAP*
RetinaNet with ResNet-34 backbone (teacher)				71.8
	✓			73.4 (+1.6)
RetinaNet with ResNet-18 backbone (student)				67.5
	✓			69.3 (+1.8)
	✓	✓		70.4 (+2.9)
	✓		✓	**72.8 (+5.3)**

TABLE 6.2 – Experimental results for thermal-only vehicle detection.

Model	Baseline	CFR	GAFF	*mAP*
Two-stream RetinaNet with ResNet-34 backbone	✓			78.0
		✓		78.9 (+0.9)
			✓	79.9 (+1.9)

TABLE 6.3 – Experimental results for multispectral vehicle detection.

6.1). Therefore, we remove the FPN neck (Lin, Dollár, et al., 2017) from the original RetinaNet and perform single-scale object detection, i.e., the RetinaNet detection head is directly attached to the output of the ResNet backbone network. To cope with multispectral inputs, we adopt the conventional two-stream network architecture to conduct feature-level fusion, where the output feature maps of two dedicated ResNet backbone networks are fused (in different ways) and sent to the joint RetinaNet detection head.

Experimental results. Table 6.1 and 6.2 report the detection performance of our trained models on monospectral images. Note that here only the visible or the thermal aerial images is used as the model input. It can be observed that our Mutual Guidance (MG) (Section 3.2) brings consistent improvements for different backbone networks and different modalities (1.5% in average). When the knowledge distillation is involved, our proposed PDF-Distil (Section 3.3) significantly outperforms the traditional feature distillation method Fitnets (Romero et al., 2015) by around 4%, and pushes the precision to 78.8% (72.8%) for visible-only (thermal-only) detection, which are more than 5% better than the baseline results from the "vanilla"

RetinaNet. Table 6.3 reports the detection results with multispectral inputs. Specifically, we evaluate the effectiveness of Cyclic Fuse-and-Refine (CFR) (Section 4.1) and Guided Attentive Feature Fusion (GAFF) (Section 4.4). The baseline multispectral fusion is the usual averaging operation between visible and thermal features. Both CFR and GAFF improve the baseline results to some extent.

6.3 Perspectives

Intelligent video surveillance will gain a lot of attention in the future world. This thesis demonstrates that multispectral scene analysis with deep learning is a powerful tool for the automatic detection and identification of objects of interest for surveillance purpose. We believe that it is a successful proof of concept for the development of more efficient and automated remote surveillance systems. Deep learning for vision is a vibrant research field, in which plenty of new technologies, network architectures and training strategies are being investigated and explored every day. In the following, we summarize what we think will be the future research trends in our studied fields:

1. For general object detection: in the short term, the backbone-neck-head architecture might continue to dominate the CNN-based object detection models (since 2017), but the room for further improvements is becoming more and more limited. Integrating **optimal transport theory** into the label-assignment strategy for object detection can be a good research direction. In the long term, we suggest reconsidering **the distinction between object detection and patch recognition**. For either two-stage, one-stage, anchor-based or anchor-free detectors, we are in fact training CNN models to recognize numerous "samples", e.g, region candidates, anchor boxes, feature points, etc. As far as the object detection task is concerned, maybe we should switch from the perspective of "samples" to "objects", e.g., the detector should recognize the actual "objects", so no redundant box should be produced and there is no need for the following rigid NMS operation. Some recent works on end-to-end object detection (P. Sun et al., 2021; J. Wang et al., 2021) and transformer-based object detection (Carion et al., 2020; X. Zhu et al., 2020) are pioneers on this field.

2. For multispectral fusion: apart from attention-based fusion, the **uncertainty** maybe also an important factor for consideration, especially when dealing with decision-level fusions. The key point about multi-sensor uncertainty might be the alignment between uncertainties from different sensors, i.e., except the absolute uncertainty estimation for each separate task or sensor, the relative uncertainty estimation inter-sensor should also be taken into account.

3. For multispectral image pair alignment: when dealing with spatially misaligned visible and thermal images, using convolution-based fusion modules might be inefficient. In this case, **transformer-based fusion modules** have global receptive-field, thus are capable to catch long-range attention and identify the correspondence between misaligned modalities.

4. For multi-sensor fusion: we plan to extend our fusion methods for scene analysis with even more modalities such as depth sensor, Doppler radar, LiDAR, etc. Evidently, how to combine these heterogeneous information (2D image, 3D point clouds, sound, etc) into a joint fusion model is the major challenge.

5. For active learning: the core of active learning methods is to set metrics to select the most informative samples for annotation. In our proposed method, the necessity of annotation is measured via the prediction inconsistency from different sensors. However, for most industrial datasets, training images are taken from video clips, and our method does not consider the relationship between adjacent frames. In fact, one could explore the **temporal prediction inconsistency** as a selection metric, e.g., if the predicted location/size/shape/category of a certain object changes drastically in a short video clip, then there might exist a detection failure for this object. Moreover, the **sampling diversity** can also be explored as a selection metric, where similar images from the same video clip should not be selected at the same time. How to measure this similarity between samples is an interesting research topic.

6. For temporal information fusion: the imaging quality of random objects from a single image may be affected by camera defocus, partial occlusion, motion blur, crowded instances, background confusion, rare poses, and other degrading factors, leading to potential detection failures. In these cases, the temporal information from the neighbouring video frames may help to better identify objects. However, most video object detection or video semantic segmentation methods resort to annotations on complete video clips, which is expensive and inefficient. From a practical point of view, we believe that **a general training framework for video object detection models based on image-based datasets** is more meaningful, in which a transformer-like network architecture (Vaswani et al., 2017; X. Wang et al., 2018) maybe helpful to encode the global historical information.

Contents

List of Figures

List of Tables

List of Abbreviations

AL	**A**ctive **L**earning
CFR	**C**yclic **F**use-and-**R**efine
COCO	**C**ommon **O**bjects in **CO**ntext
CNN	**C**onvolutional **N**eural **N**etwork
DL	**D**eep **L**earning
FPN	**F**eature **P**yramid **N**etwork
IoU	**I**ntersection-**o**ver-**U**nion
FPS	**F**rame **P**er **S**econd
GAFF	**G**uided **A**ttentive **F**eature **F**usion
MG	**M**utual **G**uidance
PDF-Distil	**P**rediction **D**isagreement aware **F**eature **Distil**lation
KD	**K**nowledge **D**istillation
MD	**M**odality **D**istillation
MR	**M**iss **R**ate
mAP	**m**ean **A**verage **P**recision
VOC	**V**isual **O**bject **C**lasses
MG	**M**utual **G**uidance
MSE	**M**ean **S**quared **E**rror
NMS	**N**on **M**aximum **S**uppression
VEDAI	**VE**hicle **D**etection in **A**erial **I**magery
ReLU	**Re**ctified **L**inear **U**nits
SOTA	**S**tate-**O**f-**T**he-**A**rt
SIAMM	**M**ulti-**M**odal **A**utomatic **I**dentification **S**ystem (**S**ystème d'**I**dentification **A**utomatique **M**ulti**M**odal)

Bibliography

Aghdam, H. H., Gonzalez-Garcia, A., Weijer, J. v. d., & López, A. M., (2019), Active learning for deep detection neural networks, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3672–3680.

Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M., (2020), Yolov4: optimal speed and accuracy of object detection, *arXiv preprint arXiv:2004.10934*.

Brazil, G., Yin, X., & Liu, X., (2017), Illuminating pedestrians via simultaneous detection & segmentation, *Proceedings of the IEEE International Conference on Computer Vision*, 4950–4959.

Brust, C.-A., Käding, C., & Denzler, J., (2018), Active learning for deep object detection, *arXiv preprint arXiv:1809.09875*.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O., (2020), Nuscenes: a multimodal dataset for autonomous driving, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Cai, Z., & Vasconcelos, N., (2018), Cascade r-cnn: delving into high quality object detection, *CVPR*, 6154–6162.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S., (2020), End-to-end object detection with transformers, *European Conference on Computer Vision*, 213–229.

Casanova, A., Pinheiro, P. O., Rostamzadeh, N., & Pal, C. J., (2020), Reinforced active learning for image segmentation, *arXiv preprint arXiv:2002.06583*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B., (2016), The cityscapes dataset for semantic urban scene understanding, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., & Zhang, L., (2021), Dynamic head: unifying object detection heads with attentions, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7373–7382.

Dalal, N., & Triggs, B., (2005), Histograms of oriented gradients for human detection, *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, *1*, 886–893.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L., (2009), Imagenet: a large-scale hierarchical image database, *CVPR*, 248–255.

Dice, L. R., (1945), Measures of the amount of ecologic association between species, *Ecology*, *26*(3), 297–302.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J., (2021), Repvgg: making vgg-style convnets great again, *arXiv preprint arXiv:2101.03697*.

Dollar, P., Wojek, C., Schiele, B., & Perona, P., (2011), Pedestrian detection: an evaluation of the state of the art, *IEEE transactions on pattern analysis and machine intelligence*, *34*(4), 743–761.

Dollár, P., Appel, R., Belongie, S., & Perona, P., (2014), Fast feature pyramids for object detection, *IEEE transactions on pattern analysis and machine intelligence*, *36*(8), 1532–1545.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q., (2019), Centernet: keypoint triplets for object detection, *ICCV*, 6569–6578.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A., (2010), The pascal visual object classes (VOC) challenge, *International Journal of Computer Vision*, *88*(2), 303–338.

Feng, C., Zhong, Y., Gao, Y., Scott, M. R., & Huang, W., (2021), Tood: task-aligned one-stage object detection, *arXiv preprint arXiv:2108.07755*.

Gao, Z., Wang, L., & Wu, G., (2021), Mutual supervision for dense object detection, *arXiv preprint arXiv:2109.05986*.

Ge, Z., Liu, S., Li, Z., Yoshie, O., & Sun, J., (2021), Ota: optimal transport assignment for object detection, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 303–312.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R., (2013), Vision meets robotics: the kitti dataset, *The International Journal of Robotics Research*, *32*(11), 1231–1237.

Girshick, R., (2015), Fast r-cnn, *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., & He, K., (2017), Accurate, large minibatch sgd: training imagenet in 1 hour, *arXiv preprint arXiv:1706.02677*.

Guan, D., Cao, Y., Yang, J., Cao, Y., & Yang, M. Y., (2019), Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection, *Information Fusion*, *50*, 148–157.

Guo, J., Han, K., Wang, Y., Wu, H., Chen, X., Xu, C., & Xu, C., (2021), Distilling object detectors via decoupled features, *arXiv preprint arXiv:2103.14475*.

Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., & Harada, T., (2017), Mfnet: towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes, *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5108–5115.

Han, S., Mao, H., & Dally, W. J., (2015), Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding, *arXiv preprint arXiv:1510.00149*.

Hazirbas, C., Ma, L., Domokos, C., & Cremers, D., (2016), Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture, *Asian conference on computer vision*, 213–228.

He, K., Zhang, X., Ren, S., & Sun, J., (2015), Delving deep into rectifiers: surpassing human-level performance on imagenet classification, *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

He, K., Zhang, X., Ren, S., & Sun, J., (2016), Deep residual learning for image recognition, *CVPR*, 770–778.

Hinton, G., Vinyals, O., & Dean, J., (2015), Distilling the knowledge in a neural network, *NIPS Deep Learning and Representation Learning Workshop*.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al., (2019), Searching for mobilenetv3, *ICCV*, 1314–1324.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H., (2017), Mobilenets: efficient convolutional neural networks for mobile vision applications, *arXiv preprint arXiv:1704.04861*.

Hwang, S., Park, J., Kim, N., Choi, Y., & So Kweon, I., (2015), Multispectral pedestrian detection: benchmark dataset and baseline, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1037–1045.

Ioffe, S., & Szegedy, C., (2015), Batch normalization: accelerating deep network training by reducing internal covariate shift, *International conference on machine learning*, 448–456.

Kao, C.-C., Lee, T.-Y., Sen, P., & Liu, M.-Y., (2018), Localization-aware active learning for object detection, *Asian Conference on Computer Vision*, 506–522.

Kim, K., & Lee, H. S., (2020), Probabilistic anchor assignment with iou prediction for object detection, *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 355–371.

Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., & Teutsch, M., (2017), Fully convolutional region proposal networks for multispectral person detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 49–56.

Law, H., & Deng, J., (2018), Cornernet: detecting objects as paired keypoints, *ECCV*, 734–750.

Leal-Taixé, L., Milan, A., Reid, I., Roth, S., & Schindler, K., (2015), Motchallenge 2015: towards a benchmark for multi-target tracking, *arXiv preprint arXiv:1504.01942*.

LeCun, Y., Bengio, Y., & Hinton, G., (2015), Deep learning, *nature*, *521*(7553), 436–444.

Li, C., Song, D., Tong, R., & Tang, M., (2018), Multispectral pedestrian detection via simultaneous detection and segmentation, *British Machine Vision Conference (BMVC)*.

Li, C., Song, D., Tong, R., & Tang, M., (2019), Illumination-aware faster r-cnn for robust multispectral pedestrian detection, *Pattern Recognition*, *85*, 161–171.

Li, H., Wu, Z., Zhu, C., Xiong, C., Socher, R., & Davis, L. S., (2020), Learning from noisy anchors for one-stage object detection, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10588–10597.

Li, Z., & Zhou, F., (2017), Fssd: feature fusion single shot multibox detector, *arXiv preprint arXiv:1712.00960*.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S., (2017), Feature pyramid networks for object detection, *CVPR*, 2117–2125.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P., (2017), Focal loss for dense object detection, *ICCV*, 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L., (2014), Microsoft COCO: common objects in context, *ECCV*, 740–755.

Liu, J., Zhang, S., Wang, S., & Metaxas, D. N., (2016), Multispectral deep neural networks for pedestrian detection, *arXiv preprint arXiv:1611.02644*.

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J., (2018), Path aggregation network for instance segmentation, *CVPR*, 8759–8768.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C., (2016), SSD: single shot multibox detector, *ECCV*, 21–37.

Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., & Zhang, C., (2017), Learning efficient convolutional networks through network slimming, *ICCV*, 2736–2744.

Loshchilov, I., & Hutter, F., (2016), Sgdr: stochastic gradient descent with warm restarts, *arXiv preprint arXiv:1608.03983*.

Lowe, D. G., (1999), Object recognition from local scale-invariant features, *Proceedings of the seventh IEEE international conference on computer vision*, *2*, 1150–1157.

Ma, N., Zhang, X., Zheng, H.-T., & Sun, J., (2018), Shufflenet v2: practical guidelines for efficient cnn architecture design, *ECCV*, 116–131.

Ma, Y., Liu, S., Li, Z., & Sun, J., (2021), Iqdet: instance-wise quality distribution sampling for object detection, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1717–1725.

Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., & Lin, D., (2019), Libra R-CNN: towards balanced learning for object detection, *CVPR*, 821–830.

Qi, L., Kuen, J., Gu, J., Lin, Z., Wang, Y., Chen, Y., Li, Y., & Jia, J., (2021), Multi-scale aligned distillation for low-resolution detection, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14443–14453.

Qin, Z., Li, Z., Zhang, Z., Bao, Y., Yu, G., Peng, Y., & Sun, J., (2019), Thundernet: towards real-time generic object detection on mobile devices, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6718–6727.

Razakarivony, S., & Jurie, F., (2016), Vehicle detection in aerial imagery: a small target detection benchmark, *Journal of Visual Communication and Image Representation*, *34*, 187–203.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A., (2016), You only look once: unified, real-time object detection, *CVPR*, 779–788.

Redmon, J., & Farhadi, A., (2017), Yolo9000: better, faster, stronger, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271.

Redmon, J., & Farhadi, A., (2018), Yolov3: an incremental improvement, *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R., & Sun, J., (2016), Faster R-CNN: towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S., (2019), Generalized intersection over union: a metric and a loss for bounding box regression, *CVPR*, 658–666.

Romero, A., Kahou, S. E., Montréal, P., Bengio, Y., Montréal, U. D., Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y., (2015), Fitnets: hints for thin deep nets, *ICLR*.

Roy, S., Unmesh, A., & Namboodiri, V. P., (2018), Deep active learning for object detection., *BMVC*, *362*, 91.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C., (2018), Mobilenetv2: inverted residuals and linear bottlenecks, *CVPR*, 4510–4520.

Siddiqui, Y., Valentin, J., & Nießner, M., (2020), Viewal: active learning with viewpoint entropy for semantic segmentation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9433–9443.

Simonyan, K., & Zisserman, A., (2015), Very deep convolutional networks for large-scale image recognition, *ICLR*.

Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al., (2021), Sparse r-cnn: end-to-end object detection with learnable proposals, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14454–14463.

Sun, Y., Zuo, W., & Liu, M., (2019), Rtfnet: rgb-thermal fusion network for semantic segmentation of urban scenes, *IEEE Robotics and Automation Letters*, *4*(3), 2576–2583.

Tan, M., Pang, R., & Le, Q. V., (2020), Efficientdet: scalable and efficient object detection, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10781–10790.

Tian, Z., Shen, C., Chen, H., & He, T., (2019), FCOS: fully convolutional one-stage object detection, *ICCV*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I., (2017), Attention is all you need, *Advances in neural information processing systems*, 5998–6008.

Viola, P., & Jones, M., (2001), Rapid object detection using a boosted cascade of simple features, *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, *1*, I–I.

Wagner, J., Fischer, V., Herman, M., & Behnke, S., (2016), Multispectral pedestrian detection using deep fusion convolutional neural networks., *ESANN*, *587*, 509–514.

Wang, J., Song, L., Li, Z., Sun, H., Sun, J., & Zheng, N., (2021), End-to-end object detection with fully convolutional network, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15849–15858.

Wang, T., Yuan, L., Zhang, X., & Feng, J., (2019), Distilling object detectors with fine-grained feature imitation, *CVPR*, 4933–4942.

Wang, X., Girshick, R., Gupta, A., & He, K., (2018), Non-local neural networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K., (2017), Aggregated residual transformations for deep neural networks, *CVPR*, 1492–1500.

Zagoruyko, S., & Komodakis, N., (2017), Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer, *ICLR*.

Zhang, H., Fromont, E., Lefevre, S., & Avignon, B., (2021a), Guided attentive feature fusion for multispectral pedestrian detection, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Zhang, H., Fromont, E., Lefevre, S., & Avignon, B., (2022), Low cost multispectral scene analysis with modality distillation, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Zhang, H., Fromont, E., Lefevre, S., & Avignon, B., (2020), Multispectral fusion for object detection with cyclic fuse-and-refine blocks, *2020 IEEE International Conference on Image Processing (ICIP)*, 276–280.

Zhang, H., Fromont, E., Lefevre, S., & Avignon, B., (2021b), Deep active learning from multispectral data through cross-modality prediction inconsistency, *ICIP: The 28th IEEE International Conference on Image Processing*.

Zhang, H., Fromont, E., Lefèvre, S., & Avignon, B., (2020), Localize to classify and classify to localize: mutual guidance in object detection, *Proceedings of the Asian Conference on Computer Vision*.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D., (2018), Mixup: beyond empirical risk minimization, *ICLR*.

Zhang, L., Liu, Z., Chen, X., & Yang, X., (2019), The cross-modality disparity problem in multispectral pedestrian detection, *arXiv preprint arXiv:1901.02645*.

Zhang, L., Liu, Z., Zhang, S., Yang, X., Qiao, H., Huang, K., & Hussain, A., (2019), Cross-modality interactive attention network for multispectral pedestrian detection, *Information Fusion*, *50*, 20–29.

Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., & Liu, Z., (2019), Weakly aligned cross-modal learning for multispectral pedestrian detection, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5127–5137.

Zhang, S., Chi, C., Yao, Y., Lei, Z., & Li, S. Z., (2020), Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9759–9768.

Zhang, X., Zhou, X., Lin, M., & Sun, J., (2018), Shufflenet: an extremely efficient convolutional neural network for mobile devices, *CVPR*, 6848–6856.

Zhang, X., Wan, F., Liu, C., Ji, R., & Ye, Q., (2019), Freeanchor: learning to match anchors for visual object detection, *arXiv preprint arXiv:1909.02466*.

Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H., (2018), Deep mutual learning, *CVPR*, 4320–4328.

Zhang, Y., Lan, Z., Dai, Y., Zeng, F., Bai, Y., Chang, J., & Wei, Y., (2020), Prime-aware adaptive distillation, *ECCV*, 658–674.

Zhang, Z., He, T., Zhang, H., Zhang, Z., Xie, J., & Li, M., (2019), Bag of freebies for training object detection neural networks, *arXiv preprint arXiv:1902.04103*.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J., (2017), Pyramid scene parsing network, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zhou, K., Chen, L., & Cao, X., (2020), Improving multispectral pedestrian detection by addressing modality imbalance problems, *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, 787–803.

Zhou, X., Wang, D., & Krähenbühl, P., (2019), Objects as points, *arXiv preprint arXiv:1904.07850*.

Zhou, X., Zhuo, J., & Krähenbühl, P., (2019), Bottom-up object detection by grouping extreme and center points, *CVPR*, 850–859.

Zhu, B., Wang, J., Jiang, Z., Zong, F., Liu, S., Li, Z., & Sun, J., (2020), Autoassign: differentiable label assignment for dense object detection, *arXiv preprint arXiv:2007.03496*.
Zhu, C., He, Y., & Savvides, M., (2019), Feature selective anchor-free module for single-shot object detection, *CVPR*, 840–849.
Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J., (2020), Deformable detr: deformable transformers for end-to-end object detection, *arXiv preprint arXiv:2010.04159*.

UNIVERSITÉ DE RENNES 1

Titre : Détection d'objets multispectrale

Mots clés : Détection d'objets, fusion multispectrales, distillation de connaissances, l'apprentissage actif

Résumé : L'analyse automatique de scènes extérieures se basant uniquement sur des images issues de caméras RGB est parfois difficile en cas d'éclairage insuffisant ou de mauvais temps. Pour améliorer la fiabilité de la reconnaissance, les systèmes multispectraux utilisent des caméras thermiques supplémentaires et détectent les objets à partir de données multispectrales. Dans ce contexte et dans cette thèse, nous avons attaqué trois verrous principaux : (1) la détection rapide et précise d'objets d'intérêt à partir d'images ; (2) la fusion dynamique et adaptative d'informations provenant de différentes modalités ; (3) la détection d'objets multispectrale à faible coût et à faible énergie et la réduction des efforts d'annotation manuelle. En ce qui concerne le premier verrou, nous optimisons d'abord l'attribution des étiquettes de l'entraînement de la détection d'objets en introduisant une stratégie de guidage mutuel entre les tâches de classification et de localisation; nous réalisons ensuite une compression efficace des modèles de détection d'objets en incluant les désaccords de prédiction enseignant/étudiant dans le cadre d'une distillation des connaissance. En ce qui concerne le deuxième verrou, trois schémas de fusion de caractéristiques multispectrales différents sont proposés pour traiter les cas de fusion les plus difficiles où différentes caméras fournissent des informations contradictoires. Pour le troisième défi, un nouveau cadre de distillation de modalité est d'abord présenté pour aborder les contraintes matérielles et logicielles des systèmes multispectraux actuels; Ensuite, une stratégie d'apprentissage actif basée sur plusieurs capteurs est conçue pour réduire les coûts d'étiquetage lors de la construction d'ensembles de données multispectrales.

Title: Multispectral object detection

Keywords: Object detection, multispectral fusion, knowledge distillation, active learning

Abstract: Only using RGB cameras for automatic outdoor scene analysis is challenging when, for example, facing insufficient illumination or adverse weather. To improve the recognition reliability, multispectral systems add additional cameras (e.g. infra-red) and perform object detection from multispectral data. Although multispectral scene analysis with deep learning has be shown to have a great potential, there are still many open research questions and it has not been widely deployed in industrial contexts. In this thesis, we investigated three main challenges about multispectral object detection: (1) the fast and accurate detection of objects of interest from images; (2) the dynamic and adaptive fusion of information from different modalities; (3) low-cost and low-energy multispectral object detection and the reduction of its manual annotation efforts. In terms of the first challenge, we first optimize the label assignment of the object detection training with a mutual guidance strategy between the classification and localization tasks; we then realize an efficient compression of object detection models by including the teacher-student prediction disagreements in a feature-based knowledge distillation framework. With regard to the second challenge, three different multispectral feature fusion schemes are proposed to deal with the most difficult fusion cases where different cameras provide contradictory information. For the third challenge, a novel modality distillation framework is firstly presented to tackle the hardware and software constraints of current multispectral systems; then a multi-sensor-based active learning strategy is designed to reduce the labelling costs when constructing multispectral datasets.