HUMAN DETECTION IN THE WILD

by
Lei Shi

A dissertation submitted to the Department of Computer Science,

College of Natural Sciences and Mathematics

in partial fulfillment of the requirements for the degree of

Doctor of Philosphy
in Computer Science

Chair of Committee: Ioannis A. Kakadiaris

Committee Member: Christoph F. Eick

Committee Member: Edgar Gabriel

Committee Member: Saurabh Prasad

University of Houston
August 2020

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Ioannis A. Kakadiaris for the continual support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this proposal. I could not have imagined having a better advisor and mentor for my Ph.D. study. Besides my advisor, I would like to thank the rest of my committee: Prof. Gabriel, Prof. Eick, and Prof. Saurabh, for their insightful comments and encouragement, but also for the questions which motivated me to widen my research from various perspectives. I thank my fellow labmates for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four years. In particular, I am grateful to Dr. Xu for enlightening me on the first glance of research. Last but not least, I would like to thank my parents for supporting me spiritually throughout writing this dissertation and my life in general.

ABSTRACT

Human detection remains a challenging task due to the problems caused by occlusion variance. Visible-body bounding boxes are typically used as an extra supervision signal to improve the performance of human detection. However, visible-body assisted approaches produce a large number of false positives, which result from a lack of adequate and discriminative full-body contextual information. As the most discriminative features of head and human, face detection has attracted much attention. Despite the great progress that has been achieved for accurate face detection, detecting multi-scale faces, especially for small faces, remains a challenging problem. Existing approaches that tackle multi-scale face detection problem could be categorized into two-stage face detectors and single-stage face detectors. Regarding two-stage face detectors, to learn discriminative facial features at various scales, the input pyramids or multi-scale feature maps are deployed to provide more facial information for the network to learn features in various scales. However, they could increase the training difficulty and complexity of the network. Regarding single-stage face detectors, feature fusion and context aggregation have been used to enrich contextual information. However, treating reliable information and noise equally could result in much noise in the fused features at different levels. Moreover, dilated convolutions in the context aggregation module could result in the gridding artifacts problem. The goal of this dissertation is to design, develop, and evaluate human detection algorithms to solve the above problems. Three contributions made in this dissertation can be summarized as follows: (i) A decoupled visible region network for human detection was designed, developed, and evaluated to overcome the occlusion challenge. The proposed human detector improved performance from MR^{-2} of 11.24 to MR^{-2} of 10.50 when compared to Bi-box which is inspired by our work on the CityPersons dataset. (ii) A two-stage face detector was designed, developed, and evaluated to overcome the scale challenge. It improves performance by mAP of 12.1% when compared to our baseline on the WIDER FACE dataset. (iii) A single-stage face detector was designed, developed, and evaluated to overcome the scale challenge. The proposed method achieves the best performance with an mAP of 77.0% on the UFDD dataset.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

1.1 Motivation

Human detection plays a crucial role in many biometrics-related applications, including pedestrian re-identification [36], traffic safety, and autonomous driving [10]. Soft biometrics plays an important role in improving performance (e.g., height, body size [1]). Therefore, human detection plays an important role in obtaining biometric information of humans to improve the performance of a biometric system. Ever since the development of convolutional neural networks (CNNs), CNN-based human detectors have dominated benchmarks [8, 41, 66]. However, despite these recent successes, occlusion is still a challenging problem. Occlusion in human detection can be categorized as either human-human or object-human occlusion. During human-human occlusion, also known as crowd occlusion, detectors can be easily confused by the similarity of features between humans. Due to the variance of object appearance, and the lack of available reliable information, it is difficult for a detector to learn discriminative features from occluded human in object-human occlusion scenario. Visible-body bounding boxes are typically used as an extra supervision signal to improve the performance of human detection to predict the full-body. However, visible-body assisted approaches produce a large number of false positives, which result from a lack of adequate and discriminative full-body contextual information.

As a representative visible-body assisted human detector, Bi-box [71] based on Faster R-CNN [38] deployed two branches for assessing the visible-body and the full-body, respectively. Nevertheless, Bi-box's performance was constrained by inadequate contextual information from the visible-body supervision signal. During training, this constraint on the effective exchange of information between branches resulted in a reduced ability to reference features and thus to enable full-body prediction. Furthermore, Bi-box ignored the contribution of features learned by the region proposal network (RPN), which are used to extract the region of interest (RoI) features. As the output of the RPN and RoI is ultimately needed by the R-CNN stage for final predictions, if the features learned by the RPN are more discriminative, the R-CNN will perform better.

As the most discriminative features of head and human, face detection has attracted much attention. Face detection plays a fundamental role in many face-related applications, such as face alignment [52], 3D face reconstruction [9], and face recognition [53]. Although current face detectors based on deep learning have achieved high detection accuracy, the detection of multi-scale faces (especially small) is still a challenging problem when images exhibit large variations in pose, blur, and occlusion. Existing approaches that tackle multi-scale face detection problem could be categorized into two directions: (1) two-stage face detectors, (2) single-stage face detectors.

For two-stage face detectors, to learn multi-scale facial features, all state-of-the-art methods deployed multi-scale feature maps, which correspond to a multi-scale feature pyramid of the input image. The input image pyramid refers to a set of images with different resolutions, which are rescaled versions of the input image. Training a two-stage detector on multi-resolution images to obtain multi-scale features results in improved face detection performance [40]. However, training on multi-resolution images increases the training time. Moreover, learning multi-scale facial features by using multi-scale feature maps produces redundant information, and increases the workload and complexity of the network.

For single-stage face detectors, S^3FD [70] based on SSD [28] is the first face detector adapting preset anchor scale to be more suitable of receptive field size for detecting various scales of faces. Moreover, it employed feature maps at different levels to detect faces with corresponding scales. However, it ignored the power of feature fusion at different levels to learn more discriminative features. SSH [33], FANet [62], and PyramidBox [44] have shown that the combination of high-level and relatively low-level semantic features can enhance multi-scale facial detection efficiency. However, merely incorporating features at different levels could result in significant noise introduced in the fused features, as these approaches treat reliable facial and noise information equally. To aggregate contextual information, SSH employed stacked convolutions to magnify the receptive field and concatenates features with different receptive field sizes. However, using stacked dilated convolutions is more efficient than using stacked convolutions to magnify the receptive field size. DSFD [19] leveraged three groups of dilated convolutions with different numbers to aggregate

contextual information, which was inspired by RFBNet [27]. However, it ignored the issue of gridding artifacts caused by dilated convolution, which leads to local inconsistent information.

To tackle the above issues of human detection and face detection in the wild, this dissertation presents our solutions.

1.2 Goal and Objectives

Our goal is to design, develop, and evaluate multi-scale human detection algorithms in the wild. Specifically, the objectives are to:

1. Design, develop, and evaluate a human detector for 2D images to overcome occlusion challenge in the wild.

2. Design, develop, and evaluate a face detector for 2D images to overcome scale challenge in the wild. The types of the designed face detector consist of:

   (a) a two-stage face detector

   (b) a single-stage face detector

1.3 Intellectual Merit

1.3.1 Human Detection

This work focused on tackling the occlusion issue of human detectors. Inspired by Bi-box [71], we proposed a network to employ additional discriminative features learned by the visible-body and the head supervision signals to enrich the contextual information of the full-body to predict the full-body. The visible-body and full-body information are both local and global when compared to each other. In addition, the head can provide stable and discriminative information for the network to predict full-body in terms of statistics related to occlusion patterns. Therefore, features learned by the visible-body and the head supervision signals could be used to enrich the contextual information of the full-body and to help the network to focus on more reliable local information.

Accordingly, to obtain high discriminative contextual information of the full-body, the attention-based feature interleaver module (AFIM) was proposed. By applying a self-attention mechanism to the RoI features and fusing attention-based RoI features, the network employed reliable contextual information for full-body prediction. Furthermore, to enhance the power of feature representation, the binary mask learning module (BMLM) was proposed, which learned binary masks of the visible-body and the full-body. Inspired by the statistics results of occlusion patterns [8], a head-aware feature enhancement module (HFEM) was proposed to provide stable and discriminative information for the network to learn discriminative features by using a head supervision signal and a supervised attention mechanism.

Our contributions, detailed in this work, include:

- An attention interleaver module that enabled highly discriminative contextual information flow between branches for predicting the visible-body and full-body.
- A binary mask learning module, which enhanced the power of represented features in the network and contributes to an improved prediction quality. These pixel-wise dense predictions enhance the sensitivity of the network for object positions and offer another source of discriminative features to the base network.
- A head-aware feature enhancement module that employed a supervised attention mechanism and a head supervision signal to provide stable and discriminative information for the network to learn discriminative features.

1.3.2 Face Detection

1.3.2.1 Two-stage Face Detector

This work focused on solving the scale problem of two-stage face detectors. Current face detectors usually use shallow and deep layers to localize small and large faces, respectively. However, features learned by different layers are pattern-variant due to the single feature pattern of the feature map. Therefore, the features used to predict multi-scale faces from different layers are not scale-invariant.

Yosinski *et al.* [58] demonstrated that deep layers are more sensitive to small input changes than shallow layers. Their work indicated that features learned by a deep CNN layer include more semantic information than the shallow layer and have a lower resolution. As for small faces, due to the existence of down-sampling operations, the loss of face resolution is exacerbated by the increase of the depth of the network. Although the shallow layer learns (with more noise) more details about an object, it does not always contribute more than a deep layer to localizing small faces in an image. Therefore, we hypothesized that using a feature map from only a deep layer is adequate to detect multi-scale faces. In this work, we conducted a comparative study to examine our hypothesis by exploring the performance impact of selecting layers with different strides and their combination.

Next, we explored the impact of contextual information based on the receptive field which represents the perceptual region for neurons in different positions of the network [31]. Neurons can't perceive all the information on the input image because of the local connection of two neighboring layers through convolution and pooling operations. If the receptive field size becomes larger, the perceptual region of neurons on input also becomes larger, which means that neurons can learn richer and more semantic-aware features. When the receptive field size becomes smaller, the learned features tend to identify local information and details. Therefore, magnifying the receptive field size benefits extracting global facial features. To enrich the contextual information, two dilated convolutions with different dilation rates were deployed to extract features from different receptive field sizes and concatenates these two features, which were followed by two convolutions to learn inter-channel features. The main contributions of this work are:

- A simple network structure that was designed to learn multi-scale facial features by employing a single-scale input image and a single-scale feature map.
- A light-weight contextual agglomeration module that was proposed to incorporate additional contextual information.

1.3.2.2 Single-stage Face Detector

This work focused on solving the scale problem of single-stage face detectors. In this work, we proposed a smoothed attention network for a single-stage face detector. We noted in our studies that the numerical variance of high-level semantic features is greater than that of low-level semantic features, which suggests that high-level semantic features will take up most of the noise in the merged features. Based on this observation, we designed an attention-guided feature fusion module (AFFM) to decrease the noise of the fused features, which applies an attention module to high-level semantic features and fused attention-focused features as well as low-level semantic features. Applying an attention module to high-level semantic features promotes the network to concentrate on reliable facial features and overlooks the noise in high-level semantic features. Therefore, the fused feature map includes less noise. Furthermore, an exhaustive analysis of the role of the attention mechanism on performance was conducted taking into account the channel-wise, spatial-wise attentions, and their combination. In addition, we developed the smoothed context enhancement module (SCEM) to circumvent the gridding artifacts problem, where dilated convolution is followed by one convolutional layer whose kernel size is related to the dilation rate of dilated convolution in order to relearn the relationship between completely separate sets of units obtained by dilated convolution. Compared to stacked convolutional layers, the SCEM also increased the computational effectiveness of magnifying receptive field size. The main contributions of this work are:

- An attention-guided feature fusion module that applied an attention mechanism to high-level semantic features can effectively decrease noise in the fused feature map.
- An smoothed context enhancement module that stacked alternative dilated convolution and classic convolution to enrich the contextual information and effectively solve the gridding artifacts problem.

1.4 Dissertation Outline

The remainder of the dissertation is organized as follows. The contributions in Objective 1 are introduced in Chapter 2. The contributions in Objective 2 are introduced in Chapter 3. Chapter 4 concludes the work.

1.5 Publications

1.5.1 Published Papers

1. **L. Shi**, X. Xu, and I. A. Kakadiaris. "Detecting multi-scale faces using attention-based feature fusion and smoothed context enhancement," *IEEE Transaction on Biometrics, Behavior, and Identity Science*, 2(2020), 235-244.

2. H. Le, C. Smailis, **L. Shi**, and I. A. Kakadiaris. "EDGE20: a cross spectral evaluation dataset for multiple surveillance problems," In *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, (Snowmass Village, CO, 2020), pp. 2685-2694.

3. **L. Shi**, X. Xu, and I. A. Kakadiaris. "SSFD$^+$: a robust two-stage face detector," *IEEE Transaction on Biometrics, Behavior, and Identity Science*, 1(2019), 181-191.

4. **L. Shi**, X. Xu, and I. A. Kakadiaris. "SEFD: a simple and effective single-stage face detector," In *Proceedings of IEEE International Conference on Biometrics*, (Crete, Greece, 2019), pp. 1-10.

5. **L. Shi**, X. Xu, and I. A. Kakadiaris. "SANet: smoothed attention network for single-stage face detector," In *Proceedings of IEEE International Conference on Biometrics*, (Crete, Greece, 2019), pp. 11-20.

6. **L. Shi**, X. Xu, and I. A. Kakadiaris. "SSFD: a face Detector using a single-scale feature map," In *Proceedings of IEEE International Conference on Biometrics: Theory, Applications, and Systems*, (LA, CA, 2018), pp. 1-10.

1.5.2 Papers Under Review

1. **L. Shi**, C. Livermore, and I. A. Kakadiaris. "DVRNet^{+}: decoupled visible region network for pedestrian detection," In *Proceedings of IEEE International Joint Conference on Biometrics*, (Houston, TX, 2020).

2 Human Detection

2.1 Related Work

Object Detectors: Human detectors have inherited many mainstream frameworks and ideas from object detectors. R-CNN [12] employed CNN to automatically perform feature extraction instead of using hand-crafted features. Fast R-CNN [11] designed a network to combine feature extractors and detectors. Faster R-CNN [38] proposed a region proposal network (RPN) and integrated RPN and R-CNN into an end-to-end training network, largely improving the computational efficiency of training a detection network. SSD [28], a pioneering work, is a single-stage detector that performs regression and classification tasks directly without the refinement process of R-CNN. SSD introduced multi-scale feature maps to learn discriminative features. Li *et al.* [20] employed a teacher-student instructional network to learn additional discriminative features for a single-stage detector.

Single-stage Human Detectors: Noh *et al.* [34] introduced partial confidence maps, allowing partial confidence predictions to be integrated into the updates of the final output tensors. Furthermore, their methods employed grid classifier averaging post-refinement to reduce the uncertainty generated by hard negative examples. ALFNet [29] stacked a series of predictors to evolve the default anchor boxes of a single-stage detector step by step, improving detection results while maintaining the accuracy of a two-stage detector and speed of a single-stage detector. The main drawback of the single-stage human detectors is the misalignment between preset anchors and convolutional features, which significantly hampers the performance.

Two-stage Human Detectors: Adapted Faster R-CNN [66] adapted and modified Faster R-CNN to human detection by tiling denser anchors, magnifying input images, refining feature stride, and ignoring region handling. Zhang *et al.* [69] leveraged channel-wise activations corresponding to visible-body parts, to implement network supervision informed by the body part heatmaps and their affiliated self-attention mechanism. Wang *et al.* [49] proposed a mechanism of repulsion loss, an adaptation of the principles of attraction and repulsion to isolate specific subjects in crowded scenes

via interactions between the different subject prediction boxes. OR-CNN [67] used the network to learn the visibility prediction for five parts of each proposal to extract discriminative features. For the Faster R-CNN based detectors, the network selects those proposals which have the intersection of union (IoU) values greater than 0.5 with their corresponding full-body ground truth as positive samples and then regress their parametric coordinates. The main drawback occurs when a positive sample IoU value may be small when compared with the visible-body ground truth, which will not aid the network in learning discriminative features due to the lack of adequate reliable information.

Other Human Detectors: CSP [30] is the first anchor-free human detector, formulating human detection as a center and scale prediction task. Adaptive NMS [26] designed a network to learn the threshold value of non-maximum suppression (NMS), which composes the network to learn discriminative features by using a suitable distribution of positive and negative samples. As occlusion levels increase, the centers of humans may be occluded. For anchor-based approaches, positive samples will not contain adequate reliable information on the visible-body, resulting in poor performance. Zhang *et al.* [65] investigated issues involving Faster R-CNN for human detection and introduced boosted forest [2], an adaption of decision tree boosting, to improve performance of the RPN in Faster R-CNN. TLL [43] leveraged temporal feature aggregation and Markov random field (MRF) [61] models to detect multi-scale humans. The main drawback of these approaches is they are trained step-by-step, resulting in low computational efficiency during both training and inference.

Bi-box [71], based on Faster R-CNN [38], devised two branches to predict the full-body and visible-body in the R-CNN stage. However, the main drawback is that the network can't effectively leverage the visible-body information to enrich the contextual information for full-body prediction, due to the lack of effective information flow between its two branches. Unfortunately, Bi-box ignored the capability of RPN feature representation, which has a demonstrated impact on performance.

2.2 DVRNet$^+$: Decoupled Visible Region Network for Human Detection

2.2.1 General Architecture

Fig. 2.1 depicts the architecture of DVRNet$^+$. DVRNet$^+$ was inspired by the Bi-box [71] architec-

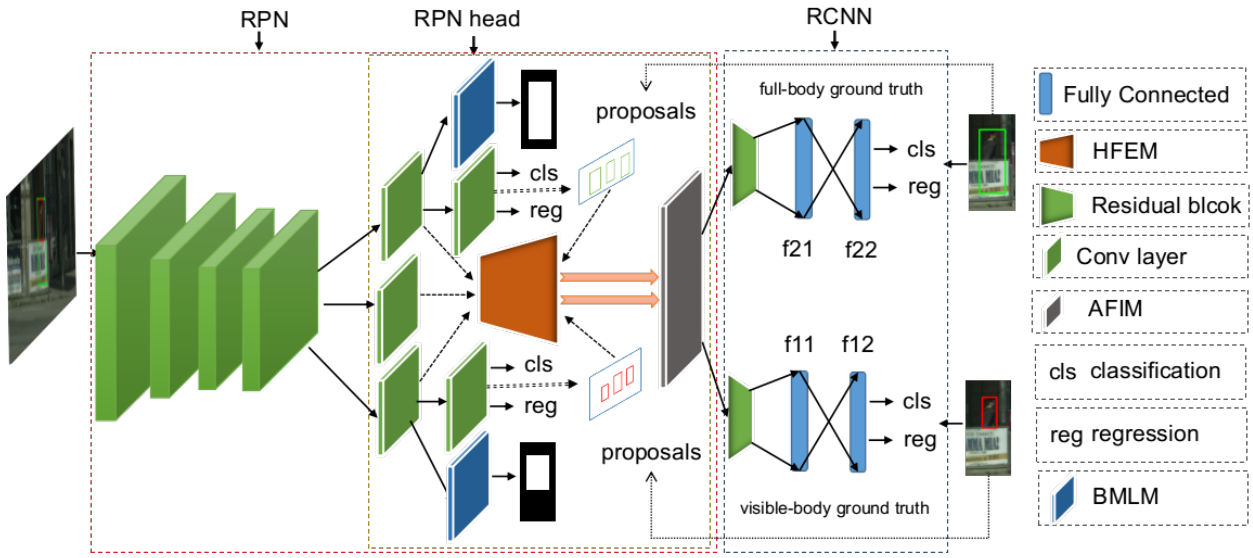


Figure 2.1: Depiction of the architecture of DVRNet$^+$. In the RPN stage, BMLM was designed to predict binary masks of the visible-body and full-body. In addition, HFEM was designed to provide stable and discriminative information for the network to learn discriminative human features that are needed by R-CNN stage. In the R-CNN stage, AFIM was proposed to enrich high discriminative contextual information provided by the visible-body supervision signal to help the network to predict the full-body. The symbols "cls" and "reg" denote the classification and regression tasks for the full-body and visible-body predictions.

ture, deploying two R-CNN branches for visible-body and full-body predictions. During training, Bi-box introduced a criterion for selecting a group of positive and negative proposals by utilizing both visible-body and full-body annotations. In this way, the RoI features that were fed into two R-CNN branches to predict the visible-body and full-body, respectively, are the same. However, the RPN of Bi-box only predicted the full-body bounding boxes. The features learned by the full-body supervision signal in the RPN were used to predict full-body and visible-body jointly in the R-CNN, which introduced the inconsistency of training in the RPN and R-CNN stages. To maintain the consistency of the training network in the RPN and R-CNN stages and richness as

well as specificity of RoI features, the RPN of our network predicted the visible-body and full-body jointly. Moreover, the visible-body and full-body ground truths were separately employed to select two groups of positive and negative samples for extracting RoI features. During testing, Bi-box added features learned by two branches in the R-CNN stage and applied the softmax function to obtain the final classification scores. However, instead of adding features, our method employed the multiplication of the classification scores obtained by these two branches as the final classification scores. Compared to Bi-box, our method resulted in high quality predictions. This modified Bi-box was dubbed mod-Bi-box and used as our baseline.

Additionally, three customized modules were proposed to help the network in full-body prediction. **Module 1:** Attention-based Feature Interleaver Module (AFIM) enriched high discriminative contextual information of the full-body to help the network to predict the full-body. **Module 2:** Binary Mask Learning Module (BMLM) performed pixel-wise classification to ensure the network focuses on more reliable details. **Module 3:** Head-aware Feature Enhancement Module (HFEM) enabled the network to learn discriminative features through stable and discriminative supervision information by a head supervision signal and a supervised attention mechanism.

2.2.1.1 Attention-based Feature Interleaver Module

Inspired by Bi-box's lack of information exchange between branches predicting visible-body and full-body, we designed AFIM to enhance the feature interaction and enrich the contextual information between two branches in the R-CNN stage. Reduced performance of detection systems in the presence of occlusions can be attributed to two challenges. The first challenge is that the appearance of the occluding object is often similar to the appearance of humans. In that case, the features learned that relate to the occluding object can distract the network easily. The second challenge is that humans are often occluded. In that case, the network can't employ adequate reliable information to predict full-body bounding boxes. Considering these two difficulties, the AFIM was designed. First, the AFIM deployed a self-attention mechanism to enhance the representation power of learned features. Although a human maybe heavily occluded, the network can still employ

discriminative features to effectively predict full-body bounding boxes. Then, the AFIM combined the attention-focused features and sent them to the branches of predicting full-body and visible-body. The fused features have the potential to enrich the contextual information of both the visible-body and full-body. When the appearance of the occluding object is similar to the appearance of the humans, the feature fusion in the AFIM has the potential to enhance the contrast of the human features and background features, which can mitigate the influence of the occluding object on the performance. When a human is occluded significantly, the attention mechanism and feature fusion can help the network to learn discriminative features and enrich contextual information of full-body. Fig. 2.2 depicts the architecture of AFIM. Specifically, the inputs of AFIM were transformed ROI

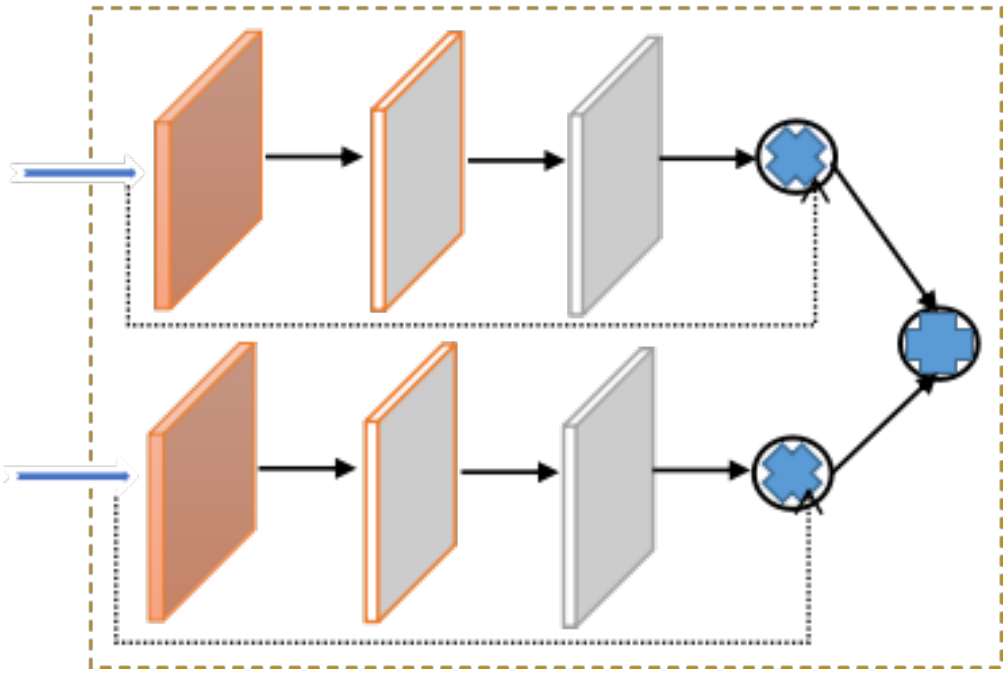
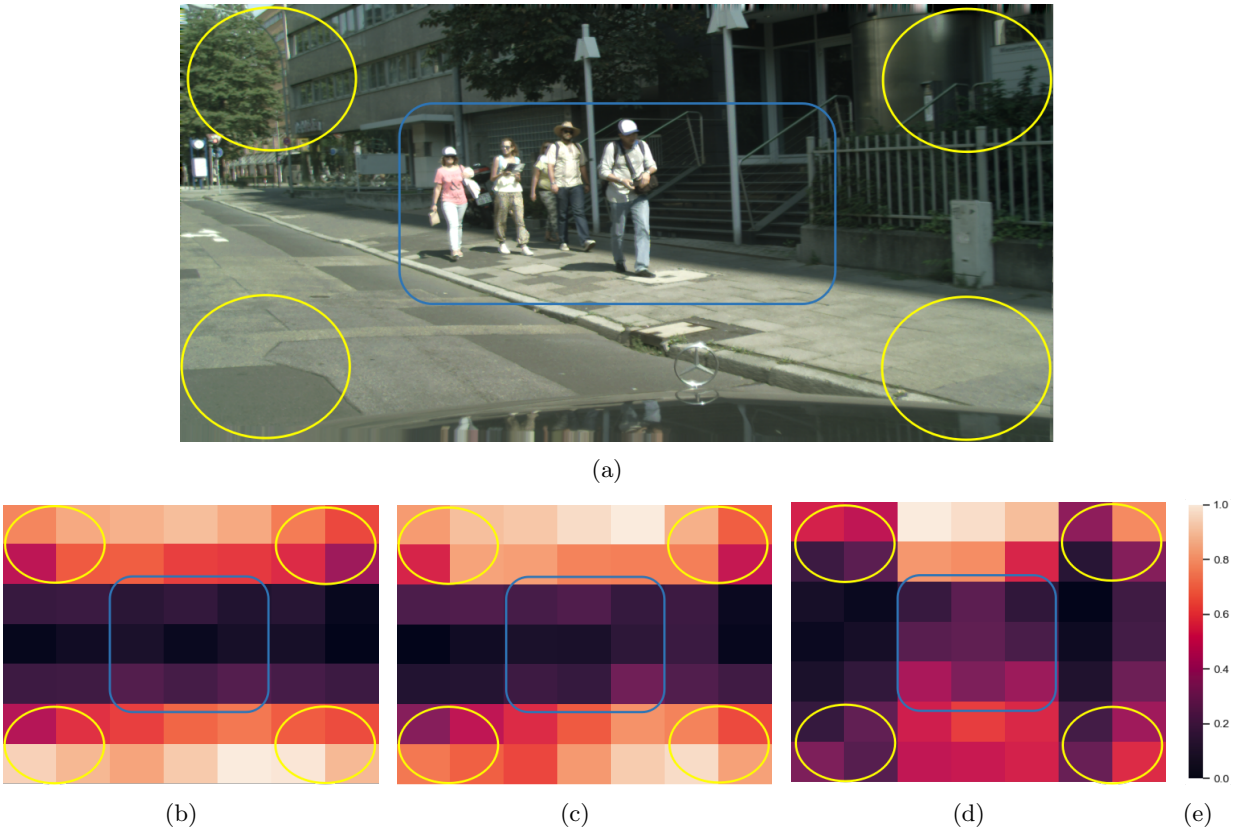


Figure 2.2: Depiction of the architecture of Attention-based Feature Interleaver Module. All convolutional layers have the same kernel size of 3×3, padding of 1, and stride of 1.

features learned by the HFEM in the RPN stage. Next, a self-attention mechanism was applied to the input RoI features to obtain the attention-based RoI features. In particular, the input RoI features were fed through three convolutional layers and a per-pixel sigmoid function to obtain the attention-based RoI features weights. The three convolutional layers have channels of 128, 64,

and 1, respectively. Moreover, the first two convolutional layers were followed by ReLU activation functions. Element-wise multiplication was applied to the input RoI features and attention-based RoI features weights to obtain attention-based RoI features. Finally, the element-wise-sum was used to fuse the attention-based RoI features from these two inputs and sent the fused features to two separate R-CNN branches, for predicting the visible and full-body, respectively. Table 2.8 and Fig. 2.3 demonstrate the effectiveness of AFIM on handling occlusion human detection.



(a) (b) (c) (d) (e)

Figure 2.3: Depiction of an input image and the heatmaps of input RoI features and fused attention-based RoI features in the AFIM. The input image has a size of 1024×2048. The features have a size of 7×7. Each cell of the heatmap corresponds to a 146×292 region in the input image. The blue rectangles include the pedestrian features; the yellow circles include the background features. (a) The input image. (b) The heatmap of the input RoI features used for predicting the full-body. (c) The heatmap of the input RoI features used for predicting the visible-body. (d) The heatmap of the fused attention-based RoI features obtained by AFIM. (e) The bar of values of the heatmaps. Comparing (b) and (c), we can observe that the values of pedestrian features are larger and that the background features are lower in (d). The observation indicates that the AFIM increases the contrast of the pedestrian features and background features. Therefore, AFIM effectively helps the network to learn more discriminative features.

2.2.1.2 Binary Mask Learning Module

The goals of the classification and regression tasks in a detection network are different. The goal of a classification task is to learn translation-invariant features. Regardless of the location of an object in a proposal, we aspire that the network classifies this proposal as a positive sample. However, the goal of the regression task is to learn translation-variant features. The expectation is that by regression the exact position of the object in the input image can be identified. We observed that the pixel-wise classification task is an effective method to enhance the sensitivity of the network at the pixel level. For a bounding box, the region-wise classification task learns global features. However, the pixel-wise classification task learns detailed features. Therefore, the pixel-wise classification and region-wise classification tasks help the network to learn additional features, enhance the sensitivity of learned features to specific pixel positions, and maintain the bounding box prediction ability. Fig. 2.4 depicts the architecture of BMLM.

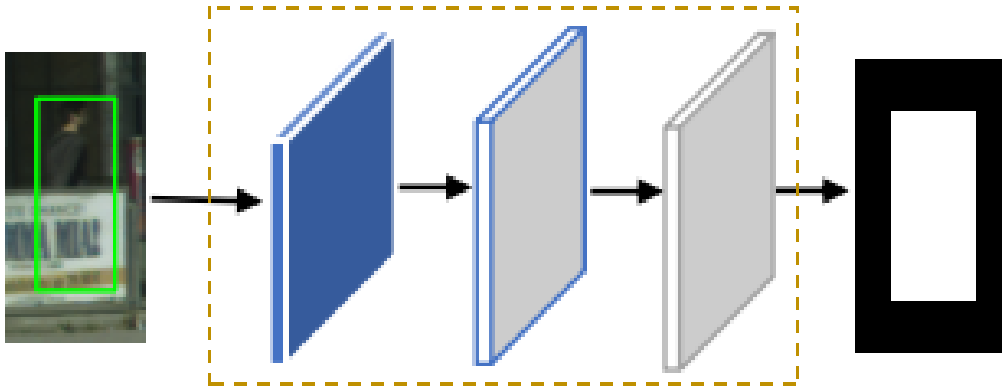
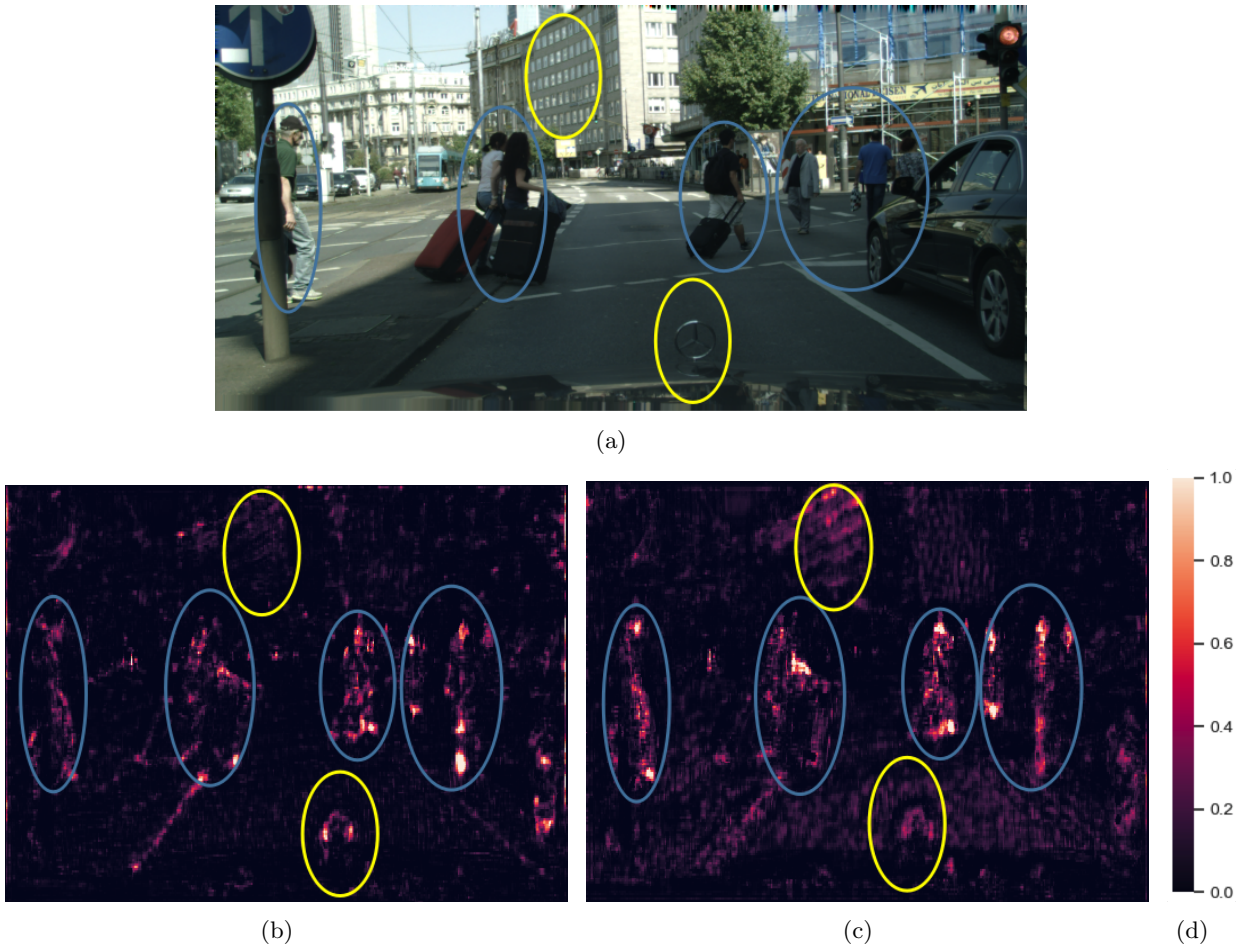


Figure 2.4: Depiction of the architecture of the Binary Mask Learning Module. All convolutional layers have the same kernel size of 3×3, padding of 1, and stride of 1.

Specifically, the BMLM performed pixel-wise classification tasks for the visible-body and full-body. The BMLM consisted of three convolutional layers with channels of 1024, 512, and 1, respectively. The first two convolutional layers were followed by Relu activation functions. Finally, a sigmoid function with cross-entropy loss was used to train the module. Table 2.8 and Fig. 2.5 demonstrate the effectiveness of BMLM on learning detailed features.



(a) (b) (c) (d)

Figure 2.5: Depiction of an input image and the corresponding heatmaps of RPN feature maps in Fig. 2.1, which are learned by RPN with and without BMLM. The input image has a size of 1024×2048. The features have a size of 128×256. Each cell of the heatmap corresponds to a 8×8 region in the input image. The blue circles include the human features; the yellow circles include the background features. (a) The input image. (b) The heatmap of the RPN features without BMLM. (c) The heatmap of the RPN features with BMLM. (d) The bar of values of the heatmaps. Comparing (a) and (b), we can observe that the human features are lighter in (c). Although some background features are shallower, and some background features are darker in (c) than that in (a) and (b), the contrast of the human features and the background features are larger in (c) than contrast in (a) and (b). The observation indicates that it is easier for the network to employ the features obtained by BMLM to distinguish the humans.

2.2.1.3 Head-aware Feature Enhancement Module

In terms of statistics related to the results of occlusion patterns [8], occlusion always happens below the head. The results also are consistent with real-world cases. With the variance of occlusion level and occluding object appearance, the information that visible-body provides to the network is uncertain because of the variance of the visible-body size and similarity of visible-body and occluding object appearance. Compared to visible-body, the head is rarely occluded. Therefore, the head could provide stable information to the network. In addition, compared to the visible-body other than head, the head appearance is more discriminative due to biometrics of faces and hair. However, the head size is far smaller than the full-body size. If the network predicts the head and full-body jointly, the network needs to tile more preset anchors with smaller scales, which will increase the workload of the network largely. Finally, the HFEM was proposed. Fig. 2.6 depicts the architecture of the HFEM.

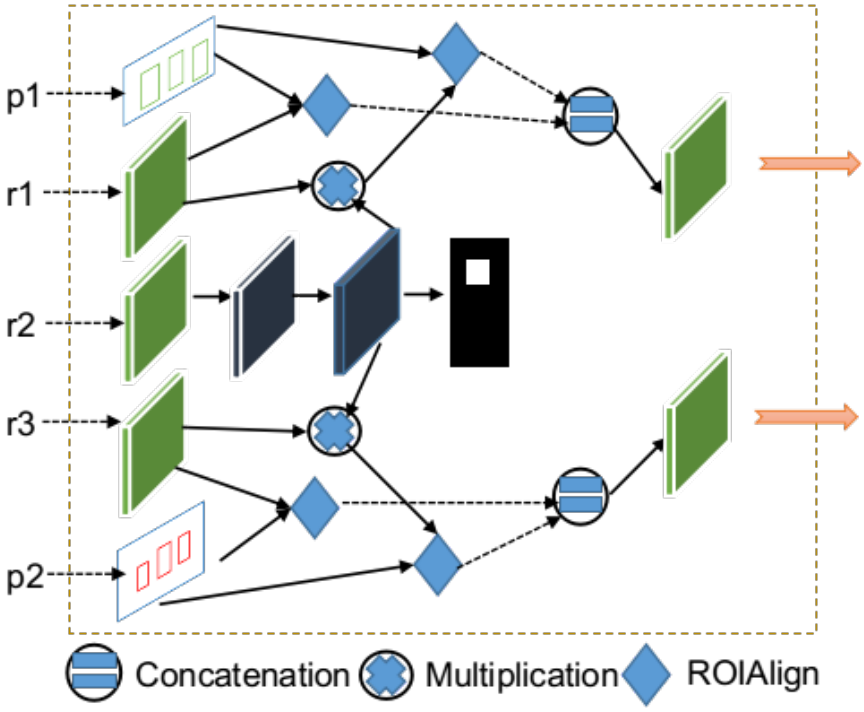
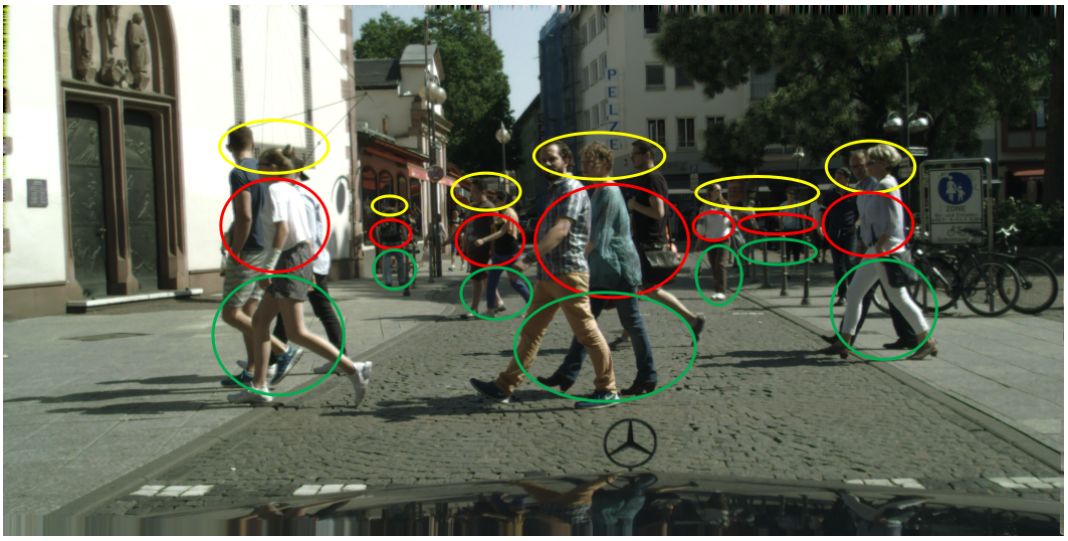
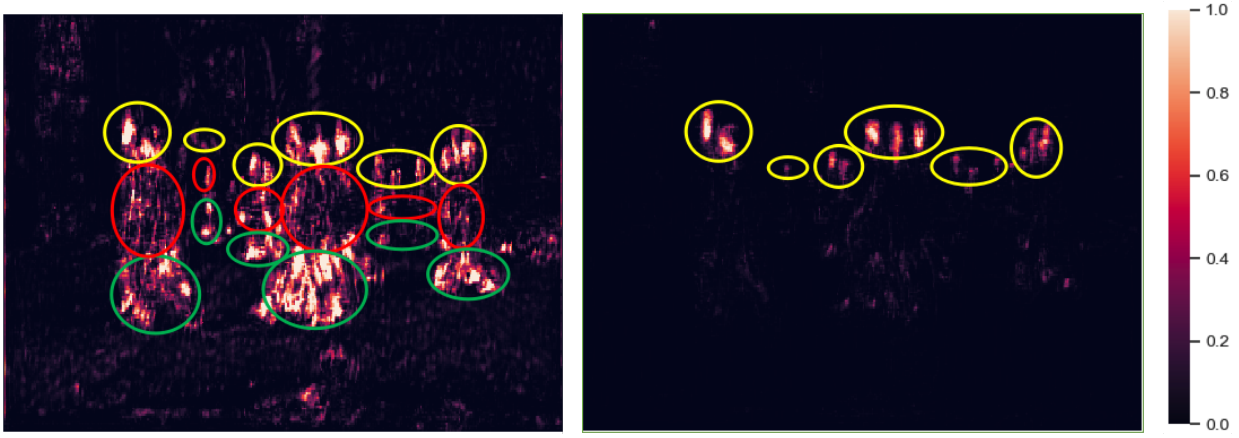


Figure 2.6: Depiction of the architecture of the Head-aware Feature Enhancement Module. All convolutional layers have the same kernel size of 3×3, padding of 1, and stride of 1.

HFEM learned a binary mask of the head by a sigmoid cross entropy loss function. The inputs of the HFEM were the features used to predict the bounding boxes and binary mask of visible-body, the features used to predict binary mask of the head, and the features used to predict the bounding boxes and binary mask of full-body in the RPN stage. The features also are denoted as r1, r2, and r3 in the Fig. 2.6, respectively. Another two inputs were the proposals selected by full-body and visible-body ground truths, respectively. They are denoted by p1 and p2 in Fig. 2.6. RoIAlign [14] was applied to the input features and selected proposals to obtain plain RoI features. Then, multiplication was applied to input features and the head binary mask to obtain supervised attention-focused features. Additionally, another RoIAlign was applied to the supervised attention-focused features and selected proposals to obtain supervised RoI features. The concatenation was applied to supervised RoI features and plain RoI features to obtain fused RoI features. Finally, one convolutional layer was used to obtain transformed RoI features from fused RoI features. The outputs were two transformed features that are fed into branches of predicting visible-body and full-body, respectively. Table 2.8 and Fig. 2.7 demonstrate the effectiveness of HFEM on helping the network to learn discriminative features.



(a)



(b) (c) (d)

Figure 2.7: Depiction of an input image and the heatmaps of RPN feature learned with and without head supervision signal. The input image has a size of 1024×2048. The features have a size of 128×256. Each cell of the heatmap corresponds to a 8×8 region in the input image. The yellow circles indicate the human features, while the red circles include the upper body features; the green circles include lower body features. (a) The input image. (b) The heatmap of the RPN features without head supervision signal. (c) The heatmap of the RPN features with the head supervision signal. Compared to (b), the features are focusing more on the head regions in (c). (d) The bar of values of heatmaps. The observation indicates that head supervision signal could provide powerful constraint for the network to learn features around head regions. Moreover, in (b), head and lower body features are more discriminative than upper body features.

2.2.2 Implementation Details

Training: Our implementation was based on the SimpleDet [6] framework with mixed precision training (FP32+FP16). ResNet-50 [15] was deployed as the backbone network. An ImageNet pre-trained model [39] was used to initialize the weights of the backbone, the layers other than the backbone network were initialized by the Xavier method [13]. The default anchor scales and ratios were set to ({4.92, 6.64, 8.97, 12.11, 16.34, 22.06, 29.79, 40.21}) and ({0.41, 1.0}) as per Zhang *et al.* [66]. The ratio parameter of 1.0 was used to smooth the total loss. In the RPN stage, IoU threshold values of 0.7 and 0.3 were used to assign anchors to positive and negative samples with a $1:1$ ratio between the two groups. The total number of positive and negative samples was 256. The threshold value of non-maximum suppression was 0.7. Before and after NMS, $12,000$ and $2,000$ anchors were retained. In the R-CNN stage, RoI-Align [14] was used to extract RoI features. The overall number of positive and negative samples at this stage was 128 maintaining the $1:1$ positive to negative ratio. The NMS threshold value of assigning anchors to positive and negative samples was 0.5. SGD [3] with a momentum of 0.9 and weight decay of 1×10^{-4} was used to train all the networks in this paper. Initially, the learning rate was set to 1.25×10^{-3}, and after $40K$ iterations, it was multiplied by 0.1. In total, there were $60K$ iterations for the CityPersons dataset. For the CityPersons dataset, training was performed on a single Tesla V100 GPU with a batch size of 1. For the CrowdHuman dataset, training was performed on two Tesla V100 GPUs with a batch size of four for each GPU, and the scales and ratios were ({32, 64, 128, 256, 512}) and ({0.41:1, 1:1, 1.5:1, 2:1, 2.5:1, 3:1}). The shorter side of the input image was 800 while the longer side of input images was not greater than $1,200$. In total, there were $450k$ iterations, the initial learning rate was 0.02, and, after $150k$ iterations, the learning rate was multiplied by 0.1.

Inference: During inference, for one full-body proposal p_f, the full-body prediction branch generated a probability score of containing a full-body, s_f, and four regression parameters of the full-body proposal, $o_f = (t_f^x, t_f^y, t_f^w, t_f^h)$ obtained by $FC2$ in Fig. 2.1; for one visible-body proposal p_v, the visible-body prediction branch predicted a probability score of containing a visible-body, s_v, and

four regression parameters of visible-body proposal, $o_v = (t_v^x, t_v^y, t_v^w, t_v^h)$ obtained by $FC4$ in Fig. 2.1. During inference, for one visible-body proposal p_v, the visible-body prediction branch predicted a probability score of containing a visible-body, s_v, and four regression parameters of visible-body proposal, $o_v = (t_v^x, t_v^y, t_v^w, t_v^h)$ obtained by $FC4$ in Fig. 2.1. Due to the confusion caused by the appearance of occluding objects, it is difficult for the network to accurately classify a bounding box. To reduce false positives, an element-wise-multiplication operation was applied to the s_f and s_v as the final probability score of containing a full-body. If one score is high and the other is low, the final score will be still low. Only if both scores are high will the final score be high. Therefore, the final score is more robust to conditions that may confuse the network. Furthermore, when the appearance of occlusion subjects is similar to the background, it is easy for the network to output the bounding boxes that extend to the background, which typically leads to the poor performance of human detection. Consequently, to mitigate this problem, the final regression parameters were set to $\lambda_1 \times o_f + \lambda_2 \times o_v$. The λ_1 was set to 0.9 while the λ_2 was set to 0.1. Through slightly modifying the output of predicting the full body by using the output of predicting the visible body, our method achieved promising performance.

Loss Function: During training, in the RPN stage, a multi-task loss that consists of bounding box classification, bounding box regression and pixel classification loss functions was defined. $L_{RPN} = \lambda_1 L_v + \lambda_2 L_f + \lambda_3 L_{BMLM} + \lambda_4 L_{HFEM}$, where L_v and L_f denotes the detection loss for predicting visible-body and full-body, respectively. Specifically, detection loss consists of classification loss and regression loss, which were identical as those defined in [11]. The symbols L_{BMLM} and L_{HFEM} denote *sigmoid* classifier with *cross-entropy* loss over two classes. In the R-CNN stage, the loss function was defined. $L_{RCNN} = \lambda_5 L_{vr} + \lambda_6 L_{fr}$, where L_{vr} and L_{fr} denotes detection loss for predicting visible-body and full-body, respectively, in the R-CNN stages. They were defined similarly as L_v and L_f. Finally, the whole network was optimized by: $L = L_{RPN} + L_{RCNN}$. The loss weights coefficients λ_1, λ_2, λ_3, λ_4, λ_5, and λ_6 were empirically set to 1.

2.2.3 Evaluation Databases and Metrics

CityPersons [66]: CityPersons dataset is a benchmark on top of the CityScapes dataset [7], which is a semantic segmentation dataset. The training set includes $2,975$ images. The validation and testing sets consist of 500 and $1,525$ images, respectively. The log-average miss rate (MR^{-2}) is used for evaluation. The dataset has four subsets. Specially, the "Reasonable" subset includes 50 pixels or taller pedestrians and occlusion equal and less than 35%; "Reasonable-Small" subset consists of 50 pixels or taller pedestrians and 75 pixels or lower pedestrians in the "Reasonable" subset. "Reasonable-Heavy" subset includes 50 pixels or taller pedestrians and 35% greater than occlusion and less than 80%; all subset consists of 20 pixels or taller pedestrians and occlusion less than 80%. In order to investigate the effectiveness of DVRNet$^+$ under different occlusion levels, we followed Zhang *et al.* [67] and Wang *et al.* [49] to split the "Reasonable" subset (occlusion equal to or less than 35%) into the "Reasonable-partial" subset (occlusion greater than 10% but less than 35%), denoted as "Partial" subset, and the "Reasonable-bare" the subset (occlusion equal to or less than 10%), denoted as "Bare" subset. For annotations whose occlusion is greater than 35% (not in the "Reasonable" set), we denoted them as "Heavy" subset.

CrowdHuman [41]: CrowdHuman is new challenging human detection benchmark. The training set includes $15,000$ images. The validation and testing sets consist of $4,370$ and $5,000$ images, respectively. In total, there are $470K$ human instances, and 22.6 persons per image with various occlusion levels. In this dataset, each human instance was annotated with a head bounding-box, a human visible-region bounding-box and a human full-body bounding-box. In addition, CrowdHuman includes more much higher crowd scenarios than other popular benchmarks. We followed the evaluation metric used in CrowdHuman [41], denoted as MR^{-2}. The full-body, visible-body, and head annotations were used for training, only full-body annotations were used for evaluation on the validation set.

EDGE20 [18]: EDGE20 is a cross spectral evaluation dataset, which includes visible and near-infrared images. The visible and near-infrared images were recorded at daytime and nighttime in

the wild, respectively. Specifically, the visible image set include $2,697$ images, the near-infrared image set consists of 797 images. It provided bounding boxes annotations of face and human. Moreover, the IDs of subjects also were saved for face recognition task. In this paper, we trained DVRNet$^+$ on the training set of CityPersons dataset and evaluated DVRNet$^+$ on the EDGE20 dataset.

2.2.4 Experimental Results

2.2.4.1 Evaluation on the Benchmarks

CityPersons: Table 2.1 summarizes the results of DVRNet$^+$'s evaluation upon the validation set of the CityPersons dataset. Compared to all selected methods, DVRNet$^+$ delivered the best results

Table 2.1: Evaluation of DVRNet$^+$ on the validation set of the CityPersons dataset. MR^{-2} was employed to compare the performance of detectors.

Methods	Reasonable	Bare	Partial	Heavy
TLL [43]	14.4	9.2	15.9	52.0
ALFNet [29]	12.0	8.4	11.4	51.9
Bi-box [71]	11.24	-	-	-
OR-CNN [67]	11.0	**5.9**	13.7	51.3
CSP [30]	11.0	7.3	10.4	**49.3**
Replusion Loss [49]	10.9	6.3	13.4	52.9
Adaptive-NMS [26]	10.8	6.2	11.4	54.0
MGAN [35]	**10.5**	-	-	-
DVRNet$^+$	**10.5**	6.59	**9.52**	55.09

as MGAN with MR^{-2} of 10.5% on the "Reasonable" subset, which demonstrated the benefit of DVRNet$^+$ for detecting occluded humans. It is worth noting that DVRNet$^+$ outperformed Bi-box by MR^{-2} of 0.74%, which inspired our work. Furthermore, DVRNet$^+$ achieved promising results with MR^{-2} of 6.59% and the best performance with MR^{-2} of 9.52% on the "Bare" and "Partial" subsets, respectively, which further demonstrated the robustness of DVRNet$^+$ to different degrees of occlusion on the "Reasonable" subset. Although DVRNet$^+$ achieved MR^{-2} of 55.09% on the "Heavy" subset, it is not strange because, as occlusion increases, the sizes of the visible-body bounding boxes can't adequately match the sizes of preset anchors defined for full-body. Therefore, a sub-optimal preset anchor setting for visible-body prediction led to the sub-optimal performance of AFIM and final predictions. This further motivated our future work on preset anchor settings while examining different sizes of ground truth at the same time.

In addition, to comprehensively compare DVRNet$^+$ with Bi-box, we evaluated DVRNet$^+$ on the "Reasonable-small" (50 pixels or taller pedestrians and 75 pixels or lower pedestrians in the

"Reasonable" subset, and it is denoted as Small(R)), "Reasonable-heavy" (50 pixels or taller pedestrians and occlusion greater than 35% but less than 80%, and it was denoted as Heavy(R)), and "All" (20 pixels or taller pedestrians and occlusion less than 80%) subsets. Table 2.2 details the results of this comparative evaluation of DVRNet$^+$ and Bi-box. We observed that under the same

Table 2.2: Comparison of DVRNet$^+$ and Bi-box on the validation set of the CityPersons dataset. MR^{-2} was employed to compare the performance of detectors. The scale denotes the scaling factor applied to the input image.

Methods	Scale	Reasonable	Small(R)	Heavy(R)	All
Bi-box [71]	1.3	11.24	47.35	44.15	43.41
mod-Bi-box	1.5	11.93	–	–	–
DVRNet$^+$	1.3	11.07	45.85	44.04	40.35
DVRNet$^+$	1.5	**10.50**	**44.22**	**43.03**	**39.92**

scaling factor of 1.3, DVRNet$^+$ outperformed Bi-box on all the subsets, which exhibited the superiority of the DVRNet$^+$ on detecting occluded pedestrians than Bi-box. With scaling factor of 1.5, DVRNet$^+$ outperformed Bi-box by MR^{-2}s of 0.74%, 3.13%, 1.12%, and 3.49% on the "Reasonable", "Reasonable-small", "Reasonable-heavy", and "All" subsets, respectively. It is worth noting that DVRNet$^+$ outperformed the Bi-box we implemented with MR^{-2} of 1.43% on the "Reasonable" subset. Therefore, our experimental results demonstrated the effectiveness and robustness of DVRNet$^+$ on detecting occluded pedestrians, compared to Bi-box.

To compare the model complexity of MGAN and DVRNet$^+$, we estimated the number of model parameters and GLOPs of MGAN and DVRNet$^+$. Table 2.3 shows the number of parameters and GFLOPs for MGAN and DVRNet$^+$. We can observe that DVRNet$^+$ has less model complexity than

Table 2.3: Comparison of DVRNet$^+$ and MGAN on the number of parameters and GFLOPs.

Methods	Parameters(M)	GFLOPs
MGAN [35]	133	15.5
DVRNet$^+$	26	3.8

MGAN, because DVRNet$^+$ has less number of parameters and GFLOPs than MGAN. Although DVRNet$^+$ has the same performance as MGAN, DVRNet$^+$ has reduced model complexity, which demonstrates the superiority of DVRNet$^+$ than MGAN.

Finally, DVRNet$^+$ was evaluated on the testing set of CityPersons dataset. Table 2.4 depicts the results of DVRNet$^+$ and compared methods. We could observe that DVRNet$^+$ achieved MR^{-2}

Table 2.4: Evaluation of DVRNet$^+$ on the testing set of the CityPersons dataset. MR^{-2} was employed to compare the performance of detectors with top results in bold.

Methods	Backbone	Reasonable
Adapted FasterRCNN [66]	VGG-16	12.97
Repulsion Loss [49]	ResNet-50	11.48
OR-CNN [67]	VGG-16	11.32
Adaptive-NMS [26]	ResNet-50	11.40
DVRNet$^+$	ResNet-50	**11.17**
MGAN [35]	VGG-16	**9.29**

of 11.17% and promising performance compared to all compared methods, which demonstrated the robustness of DVRNet$^+$ on handling occlusions.

CrowdHuman: Table 2.5 summarizes the results of our evaluation of DVRNet$^+$ and compared approaches on the CrowdHuman dataset. To better understand the impact of FPN on performance,

Table 2.5: Evaluation of DVRNet$^+$ on the validation set of the CrowdHuman dataset.

Methods	FPN [23]	MR^{-2} %
Faster R-CNN [38]		50.42
RetinaNet [24]		63.33
DVRNet$^+$	×	62.02

we re-implemented and evaluated the Adapted Faster R-CNN [66] on CityPersons and CrowdHuman. Our implementation of Adapted Faster R-CNN achieved 15.26% MR^{-2} for the input image with a scale of 1.0 on the "Reasonable" subset of the validation set of the CityPersons dataset, while 15.4% MR^{-2} of Adapted Faster R-CNN was reported in Zhang *et al.* [66]. However, on the CrowdHuman dataset, we obtained the same conclusion as Liu *et al.* [26] that it fails to be a baseline due to poor performance. After analyzing the results, we concluded that after resizing the input images, it would be very difficult for the network to learn discriminative human features in dense crowd scenarios due to the large variance of scale. DVRNet$^+$ based on Faster R-CNN didn't employ feature pyramids, yet achieves 62.02% MR^{-2}, which demonstrated the effectiveness of DVRNet$^+$ on overcoming occlusion challenge.

EDGE20: Table 2.6 and Table 2.7 summarize the results of our evaluation of DVRNet$^+$ and the Bi-box implemented by us on the VIS and NIR images, respectively.

Table 2.6: Evaluation of DVRNet$^+$ and Bi-box implemented by us on the day session of the EDGE20 dataset.

Methods	MR^{-2} %
DVRNet$^+$ [38]	18.7
mod-Bi-box	23.2

Table 2.7: Evaluation of DVRNet$^+$ and Bi-box implemented by us on the night session of the EDGE20 dataset.

Methods	MR^{-2} %
DVRNet$^+$	85.8
mod-Bi-box	100.0

Both DVRNet$^+$ and mod-Bi-box were trained on the training set of CityPersons dataset with a scale of 1.5 and evaluated on the EDGE20 dataset with the same size of training images. The observation we can get is that our method achieved better performance on both VIS and NIR image sets. Especially, DVRNet$^+$ outperformed the baseline by MR^{-2} of 4.5% and 14.2% on the VIS and NIR image sets, respectively. It is worth noting that DVRNet$^+$ was only trained on the visible images on the training set of CityPersons and evaluated on the NIR images, which suggested that the visible image and near-infrared image could share some features learned by the neural network. However, the difference between the visible image and near-infrared image should be learned for further improving the performance of DVRNet$^+$ on the near-infrared images. Therefore, how to help the network to understand the difference between visible and near-infrared images will be our further work.

2.2.4.2 Ablation Study

An ablation study was conducted to investigate the effectiveness of AFIM, BMLM, and HFEM on detecting occluded humans. Table 2.8 summarizes the results of the ablation study.

Table 2.8: Ablation study of DVRNet$^+$ on the validation set of the CityPersons dataset. The scale denotes the scaling factor applied to the input image. MR^{-2} is employed to compare the performance of detectors (lower score indicates better performance).

Methods	Scale	Reasonable	Bare	Partial	Heavy
mod-Bi-box	1.0	14.89	8.97	16.44	60.10
mod-Bi-box	1.5	11.93	7.93	12.32	57.96
mod-Bi-box + AFIM	1.5	10.84	7.03	10.07	56.69
mod-Bi-box + BMLM + AFIM	1.5	10.57	6.82	9.63	56.31
mod-Bi-box + BMLM + AFIM + HFEM	1.5	**10.50**	**6.59**	**9.52**	**55.09**

AFIM: The baseline achieved 11.93% MR^{-2} on the "Reasonable" subset, failing the Bi-box [71] by 0.69% MR^{-2}. Compared to the baseline, the AFIM reduced the MR^{-2} by 1.09% from 11.93% to 10.84% on the "Reasonable" subset. It is worth noting that the results of adding AFIM to the baseline outperformed Bi-box by 0.4% MR^{-2} on the "Reasonable" subset, which has demonstrated the superiority of the AFIM on assisting the network in handling occlusion problem by enriching discriminative contextual information. Moreover, the AFIM reduced the MR^{-2} by 0.9%, 2.25%, and 1.27% from 7.93%, 12.32%, and 57.96% to 7.03%, 10.07%, and 56.69% on the "Bare", "Partial", and "Heavy" subsets, respectively, which further demonstrated the effectiveness of the AFIM on improving the performance of detecting occluded humans. Fig. 2.3 also illustrates that the AFIM can help the network to learn more discriminative features.

BMLM: The BMLM improved the MR^{-2} from 10.84%, 7.03%, 10.07%, and 56.69% to 10.57%, 6.82%, 9.63%, and 56.31% on the "Reasonable", "Bare", "Partial", and "Heavy" subsets, respectively. The BMLM performed pixel-wise dense predictions, which could provide more details for the network to learn discriminative features. Compared to each other, the visible-body prediction was a local-wise prediction and full-body prediction was a global-wise prediction. Therefore, the RPN with BMLM helped the network to learn more discriminative features by performing pixel-wise, local-wise, and global-wise predictions. Fig. 2.5 also demonstrates this point. When designing the

architecture of the BMLM, we also considered the supervised attention mechanism. It means the input features were multiplied with the binary mask. Due to the supervision signal of the binary mask, the supervised attention mechanism often provides more robust features through increasing the contrast of the importance of reliable information and background information. However, after conducting the experiments, We were surprised to find out that the supervised attention-based BMLM failed BMLM MR^{-2} of 1.12% on the "Reasonable" subset. Through analysis, we observed that the binary mask provides unstable importance value of features to the network due to variance of occlusion level. Therefore, we hypothesized that the head binary mask could perform better than visible-body and full-body binary masks. Because the head binary mask could provide more stable information.

HFEM: The HFEM improved the MR^{-2} of 0.07%, 0.23%, 0.09%, and 1.22% on the "Reasonable", "Bare", "Partial", and "Heavy" subsets, respectively. It is worth noting that HFEM largely improved the performance under the circumstance of heavy occlusion. For better understanding the function of supervision attention mechanism on the performance, we also conducted the experiment where the HFEM precluded the supervised attention mechanism. The performance of HFEM without a supervised attention mechanism failed HFEM with a supervised attention mechanism MR^{-2} of 0.67% on the "Heavy" subset, which also demonstrated our hypothesis that the head supervision signal can provide more stable and discriminative features under the condition of heavy occlusion. Fig. 2.7 depicts the heatmaps of RPN features learned with and without HFEM. One surprising observation was that the features learned by RPN with HFEM were focusing on the head regions. Moreover, the network didn't learn discriminative features from other body parts. This observation demonstrated that our hypothesis that the head supervision signal offers more powerful control than visible-body and full-body supervision signals due to stable information provided for the network by the head regions. For the full-body supervision signal, the body features learned by the network could be influenced by the appearance of occluding objects. For the visible-body, due to the variance of the size of the visible-body, it is not easy for the network to learn stable and discriminative features from unstable information. Therefore, we could conclude that the network

tends to learn more discriminative features from the supervision signal that could provide stable supervision information. Another surprising observation was that the network tended to learn more discriminative features from the head and lower-body regions than upper-body regions when the visible-body and full-body supervision signals were used. This observation also was consistent with the perception that due to the movement of the lower-body and discriminative features from faces and hair, the head and lower-body can provide more discriminative information for the network. Compared to the head and lower-body, most of the upper-body is dominated by the appearance of clothes. Thus, the upper-body can't provide sufficient discriminative information for the network.

3 Face Detection

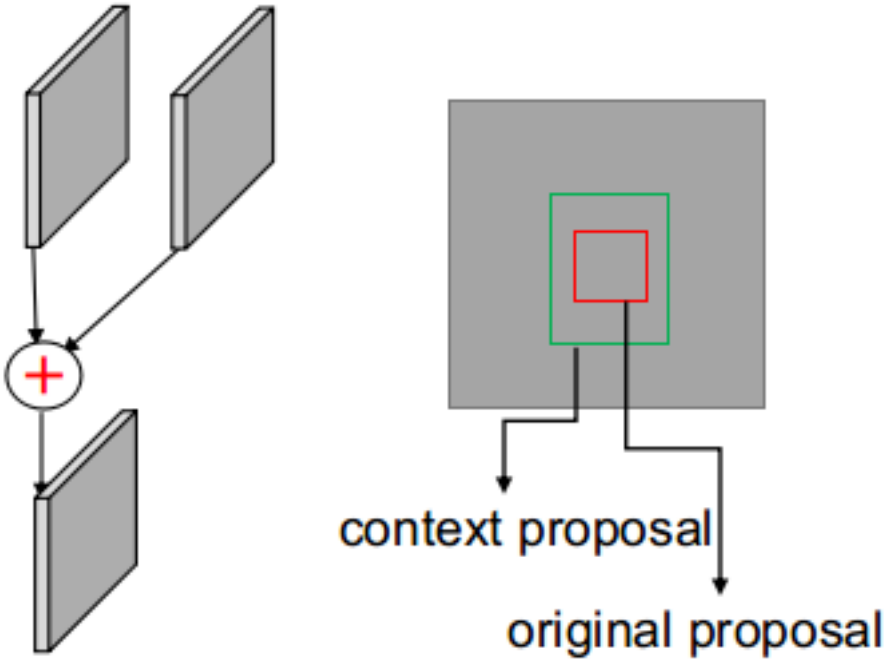
3.1 Related Work

Multi-scale Features: The simplest way of extracting multi-scale features is building an image pyramid. After resizing a single input image to form a set of images with different resolutions, the image at each resolution is fed into the network to learn facial features at a corresponding scale. The feature pyramid is a popular method to extract multi-scale features to detect faces of various scales. In general, the numbers of extra layers and the levels of the feature pyramid vary depending on the objective. Single-stage face detectors [28, 33, 44, 62, 70] used a feature pyramid as a primary way of detecting multi-scale faces; they leveraged feature maps from a backbone network and combine them as a feature pyramid with different levels to predict face locations directly without fully-connected classifier layers. Due to the existence of down-sampling layers, a network can obtain feature maps of different scales. Therefore, to cover the shortage of low-resolution feature maps in deep layers, one-stage face detectors were trained to learn face templates of different scales to detect multi-scale faces. In particular, the SSD developed by Liu *et al.* [28] and the S^3FD developed by [70] used a feature pyramid of a single level, which directly used feature maps from different layers in the backbone network to localize multi-scale faces. Faces of one scale were predicted by a feature map of a specific scale, which depended on the input image size and the stride of the feature map. To adapt the anchor's scale to that of small faces, S^3FD offered a new anchor matching strategy in which the anchor scale was set in terms of the receptive field size in the feature map and it was used to predict faces of the corresponding scale. SSH [33], PyramidBox [44], and FANet [62] introduced different strategies on feature fusion of multi-level feature pyramid.

SSH fused the feature maps of *conv3-3* and *conv4-3* of the VGG16 network in the direction of the channel, which produced a feature pyramid of two levels. Both PyramidBox and FANet used a three-level feature pyramid. Through a set of down-sampling operations on the last layer of the backbone, this produced three additional feature maps. Faster R-CNN [38] is the milestone approach for two-stage detectors.

First, bounding box candidates were being proposed by the first stage detector. Then, the second stage detector refined the region proposals using another part of the network. Face R-CNN [47] fitted Faster R-CNN to face detection and employed multi-scale training and multi-scale testing to improve the performance of detecting multi-scale faces. The method SkipFace developed by Samangouei *et al.* [40] concatenated feature maps from the backbone network to classify faces and regress bounding boxes in the first stage. Compared with SkipFace, HR [16] learned three templates, focusing on small, medium and large faces, respectively, and achieved state-of-the-art performance. Multi-scale FCN, developed by Wang *et al.* [50], allowed additional aggregation of different layers. Although additional layers or multi-scale inputs include multi-scale facial information, they also produce redundant information. Feature maps from different layers capture different semantic information. Therefore, multi-scale features learned by different layers don't have unified interpretability. Moreover, multi-scale feature maps and input pyramid increase the complexity of the network. To solve these problems, we designed a simple and effective network to learn multi-scale facial features by using a single-scale input image and a single-scale feature map.

Contextual Information: Contextual information is leveraged to detect faces in challenging conditions such as small faces, extreme pose and severe occlusions. In such conditions, very little facial information can be retrieved from the face region. With the help of contextual information (e.g., hair, ears or even body), it is easier for a network to detect faces. Common ways to increase contextual information fall into two categories. The first category includes methods that merge the feature maps. The second category includes methods that enlarge the anchor region and feed enlarged proposals into the classifier and the regressor. Fig. 3.8 depicts these two manners of increasing contextual information. HR [16], trained different templates (larger than the detection window size) and employed them to detect faces at multiple scales. The method Context classifier developed by Samangouei *et al.* [40] used two branches: the first branch followed the same design as Faster R-CNN, and the second branch refined enlarged region proposals, which was executed in parallel with the second stage of the first branch.



(a) (b)

Figure 3.8: Depiction of the two common ways to increase contextual information. (a) Feature fusion of two different feature maps (the operator "+" denotes *Element-Wise-Sum* or *Concatenate*); (b) Region enlargement (the inner box denotes the original region proposals and the outer box the enlarged region that contains additional contextual information.)

The developers demonstrated that contextual information can improve performance up to 10% when using low-resolution input in the hard set of WIDER FACE validation set. But with high-resolution input, the improvements were minor at double the computation cost. ScaleFace [57] combined two adjacent residual blocks of He *et al.* [15] to improve contextual information. The results indicated that prediction using feature maps with a stride of eight has better performance than other conditions. Meanwhile, Yang *et al.* [57] trained three branches for small, medium, and large faces, respectively. However, their algorithms didn't consider feature maps of higher resolution which can provide additional facial information, especially for tiny faces. Zhu *et al.* [72] presented three approaches to magnify contextual information through reducing stride of feature map; dilated convolution achieved the best performance. SSH [33], provided a new context module to augment contextual information. In the context module, two banks of convolution operations were applied to the same feature map: one single convolution and one with two convolution layers. Finally, their output was concatenated in the channel dimension as an input for prediction. Yu *et al.* [60] suggested that the receptive field size can be magnified by increasing the depth of the network. Therefore, the context module of SSH combined two sub-modules with different receptive field sizes, which enabled rich contextual information in this feature map. However, with the increase of the number of stacked feature maps, the computational efficiency of the network drops. To solve this problem, a computationally light-weight context agglomeration module was proposed to incorporate additional contextual information.

Other Detectors: Yolo [37] considered the detection task as a regression task and replaced bounding box proposals by dividing the whole input into small non-overlapping squares to improve inference speed. MTCNN [63], leveraged coarse-to-fine structures to train three cascaded CNNs for joint face detection and alignment, which has demonstrated that a landmark localization step can help reduce false positives in a face detection task. Li *et al.* [21] deployed a 3D model into the face detector to localize facial key points, thereby helping to improve face detection. FAN [48], introduced an attention mechanism to detect occluded faces, where the binary mask of face and background served as an input to the attention sub-network. By learning discriminative features

of occluded faces and background, FAN effectively improved the performance of occluded face detection. Cascade R-CNN [4], focused on assigning anchors to positive and negative examples in terms of Intersection of Union (IoU) between the anchor and its ground truth for the two-stage detector. The results indicated that the second stage using a cascade structure to train with the increasing of IoU threshold value can effectively reduce close false positives, which helped the network find additional discriminative facial features. Although these methods have demonstrated their effectiveness on face detection benchmarks, they increased the complexity and difficulty of training the network.

3.2 SSFD$^+$: A Robust Two-stage Face Detector

3.2.1 General Architecture

Fig. 3.9 depicts the general architecture of SSFD$^+$, which is based on the VGG16 network. SSFD$^+$ followed the work of Faster R-CNN, which consists of two stages: (i) generating face bounding box candidates according to features extracted from the backbone network, (ii) refining proposal candidates to obtain the final face classification and localization predictions. In the first stage,

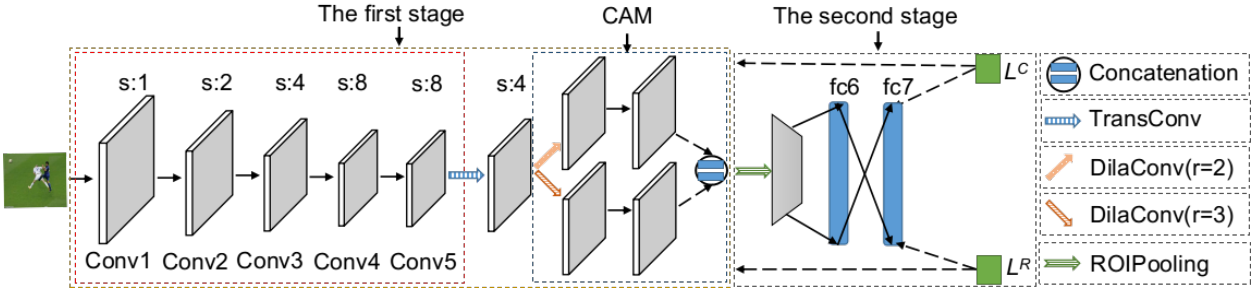


Figure 3.9: Depiction of the architecture of SSFD$^+$. The Context agglomeration module (CAM) consists of two dilated convolutions with dilation rates of 2 and 3, respectively, two regular convolutions with kernel of 1, and the concatenate operation. "TransConv" refers to the transposed convolution operation, while "DilaConv" refers to the dilated convolution operation. The symbols "L^C" and "L^R" denote the classification and regression losses, respectively.

the SSFD$^+$ backbone network included all convolution operations of VGG16 plus one context agglomeration module. In the second stage, the output of the context agglomeration module was

fed into a region-wise feature extraction layer and two fully connected layers, which were leveraged to predict classification scores and the bounding box regression offsets. An RoI-Pooling layer was used to extract the region-wise features.

3.2.1.1 Single-scale Feature Map

Our objective was to learn multi-scale facial features using a single-scale feature map and a single-scale input image instead of a feature pyramid and an input pyramid. First, the pooling operation which reduced the feature map from 8 to 16 in the backbone network has been removed. Therefore, the downsampling rate of the feature map of the last convolution layer in the original VGG16 is 8, which could preserve high-resolution information. Then, transposed convolution was deployed to amplify the last feature map with a scale of 2 in the backbone, which made the network obtain additional facial information, and tile more and denser anchors to detect multi-scale faces. The number of filters of the transposed convolution layer was set to 512.

3.2.1.2 Context Agglomeration Module

The receptive field size can be used to adjust to the abstract level of features obtained by each layer. The work of Luo *et al.* [31] indicated that the receptive field has two variants, theoretical receptive field and effective receptive field. Not every input unit contributes equally to output in the theoretical receptive field. Moreover, the contribution of inputs to the non-central region is far smaller than to the center of the theoretical receptive field. Therefore, we focused on the central area of the theoretical receptive field, which is called effective receptive field. According to Luo *et al.* [31], the effective receptive field has a Gaussian distribution and its area is about a quarter of the theoretical receptive field. The effective receptive field size has a positive proportion relationship with the semantic level of features extracted by the network. To maintain a high-resolution feature map, the greater the receptive field size is, the higher global facial information we can obtain. If the effective receptive field size is too small, the network can't localize large faces due to inadequate facial information. Moreover, without adequate contextual information,

small faces can't be detected as well. Usually, increasing the number of convolution operations can enlarge the receptive field size, however additional convolution operations also increase the number of parameters of the network. Dilated convolutions can be deployed to produce a larger receptive field size through dilation parameters of convolution kernels without increasing the number of parameters compared with regular convolutions [5, 59].

In our work, inspired by receptive field theory and methods of increasing contextual information, two dilated convolutions followed by two regular convolutions were applied to the last feature map in the backbone network to extract two feature maps with inter-channel information from different receptive field sizes. Then, the feature maps were concatenated to enrich contextual information. These two dilated convolutions have dilation rates of 2 and 3, respectively. The two convolutional layers have the same kernel size of 1×1.

3.2.2 Implementation Details

Training and Inference: VGG16 was chosen to be the backbone network, and the pre-trained model trained on ImageNet was used to initialize the weights of all networks. All input images were normalized and rescaled so that the shorter side was $1,200$ pixels, and the longer side was not greater than $1,600$ pixels. The scale setting of default anchors followed the face scale distribution of WIDER FACE training set and was set from 4 to 512 ({4, 8, 16, 32, 64, 128, 256, 512}) in Zhang *et al.* [70]. The ratio setting of default anchors was {0.5, 1, 2}. In the first stage, the IoU threshold values used to assign anchors to positive and negative samples were 0.7 and 0.3, respectively. The ratio of positive and negative samples was $1 : 1$. The total number of positive and negative samples was 256. The threshold value of non-maximum suppression was 0.7. Before and after non-maximum suppression, $12,000$ and $2,000$ anchors were retained. Batch size was set to 1 and 2 in the two stages respectively. In the second stage, RoIPooling was used to extract region-wise features of the detection layer, followed by fully connected layers of VGG16. The overall number of positive and negative samples was 128. The ratio of positive to negative samples was $1 : 3$. The threshold values of positive and negative samples were 0.5. SGD with the momentum

of 0.9 and weight decay of 5×10^{-4} was used to train our model. Finally, the learning rate was set to 0.001 initially, and after 7 epochs, it was multiplied by 0.1. In total, there are 10 epochs and every epoch includes $25,760$ iterations. The hard negative sample mining technology was applied to all baselines. and followed as the basic rule of Shrivastava *et al.* [42]. In the region proposal step, an anchor was assigned to a positive or negative sample in terms of IoU of anchors with ground truth. If the IoU is too big, it can result in a few positive samples, thereby overfitting. If the IoU is too small, many positive samples contain a lot of backgrounds, which leads to having difficulty rejecting close false positives for the detector. We have experimented with different IoU threshold values and found that 0.5 offered a good balance between performance and speed. Multi-scale testing was deployed to improve the performance, which followed the pyramid testing setting of SSH. Specifically, the shorter side of the input image was resized to 800, and the longer side of the resized image should be not greater than $1,200$. Then, the shorter side of the resized image was resized to 500, 800, $1,200$ and $1,600$. Finally, the bounding box voting strategy was deployed to produce final predictions.

Loss Function: We defined a multi-task loss function for classification and regression following the approach of Ren *et al.* [38]. For the face classification task, the *softmax* classifier uses classic *cross-entropy* loss. *Smooth* L_1 loss was employed for the regression task. Therefore, the total loss was formulated as the sum of classification and regression losses:

$$L\left(p_i, t_i\right) = \frac{1}{N} \sum_{i}^{N} L_c\left(p_i, p_i^*\right) + \frac{1}{N} \sum_{i}^{N} p_i^* L_r\left(t_i, t_i^*\right), \tag{3.1}$$

where the first term of the sum corresponds to the classification loss: i denotes the index of an anchor, and p_i denotes probability predicted for that anchor. The ground-truth label which is denoted by p_i^* is equal to 1 if anchor i is a positive sample, otherwise 0. The *softmax cross-entropy* loss over two classes of positive and negative samples was denoted by L_c. Intersection over Union (IoU) is the criterion for defining positive and negative samples. In this paper, if the IoU of a bounding box and its ground truth is larger than 0.7, the bounding box was defined as a positive

sample, whereas if its IOU is smaller than 0.3, it was defined as a negative sample. The total classification loss was normalized by $\frac{1}{N}$, where N represents the number of anchors selected to the loss. The second term of the sum is the regression loss, where t_i denotes four predicted parametric coordinates defined by Girshick *et al.* [12], and the corresponding ground truth is denoted by t_i^*. The symbol L_r denotes the *smooth* L_1 loss, which is activated by $p_i^* = 1$. Similar to the classification case, the total regression loss was normalized by $\frac{1}{N}$. In this work, N was equal to 256 and 128 for the first and second stages, respectively. Finally, the classification and regression losses were added to form the total loss.

3.2.3 Evaluation Databases and Metrics

AFW Dataset [73]: The AFW dataset contains 205 Flickr images, where each image contains at least one large face. In total, there are 473 annotated faces of large variations in both pose and scale. Moreover, this dataset provided one bounding box and six landmarks for each face.

PASCAL Faces Dataset [54]: This is a subset of PASCAL dataset including only faces. It includes 851 images and $1,341$ annotated faces with large variations of face appearance and pose. It was collected from the PASCAL person layout test subset.

FDDB Dataset [17]: It consists of $2,845$ images with $5,175$ labeled faces. FDDB faces are annotated by ellipse coordinates. Moreover, the face images are related to a wide range of difficulties including occlusions, difficult poses, and low resolution and out-of-focus faces. Furthermore, this dataset provided the specification of face regions as elliptical regions.

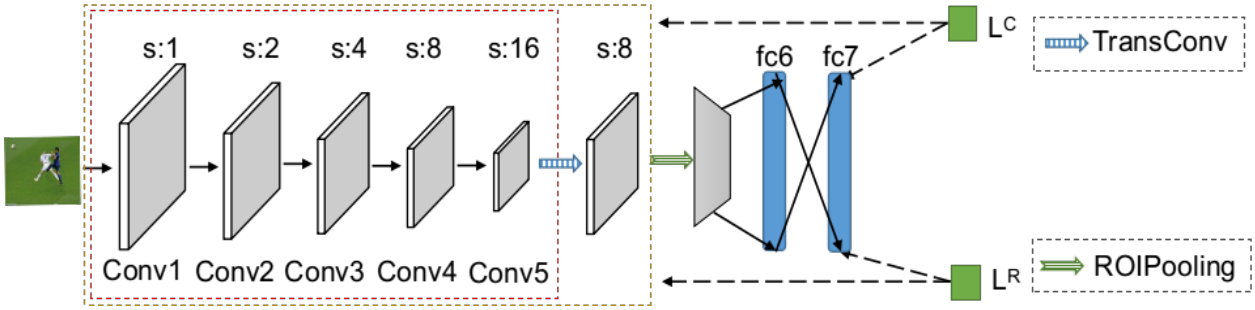
WIDER FACE Dataset [55]: It is used to train, validate and test the networks. It includes $12,880$ images as the training set, $3,033$ images as a validation set and $16,290$ images of test set. In total, the training set includes $158,989$ annotated faces, the validation set consists of $39,496$ annotated faces and the test set has $195,218$ annotated faces. The training set of all baselines and SSFD$^+$ in this paper includes the original training set of the WIDER FACE training set and their horizontal flip augmentation. All baselines were validated on the original validation set.

3.2.4 Experimental Results

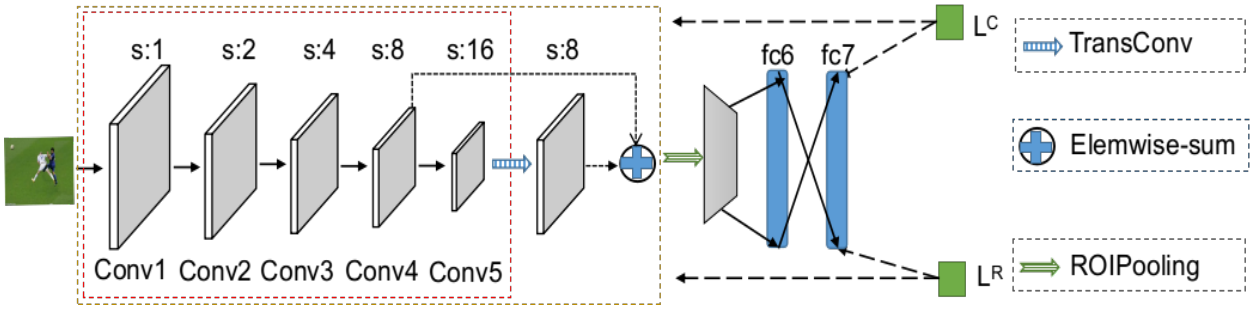
3.2.4.1 Design Consideration

Baseline Networks: In the first stage of our baselines, all convolution operations of VGG16 network were leveraged to generate face proposals. A detection layer followed by two convolution layers was deployed to predict classification scores and face regions from feature maps with strides of 4, 8 and 16 (named F_4, F_8 and F_{16}), which are obtained from the original *conv3*, *conv4*, *conv5* layers of VGG16, respectively. The feature maps obtained by applying transposed convolution operation (TransConv) with a scale of 2 on F_{16} and F_8 were denoted as T_{16} and T_8 respectively. Fig. 3.10 depicts three baselines based on Faster R-CNN. To explore the effect of feature maps from different layers and their combination on detecting multi-scale faces, we performed comparative experiments with baselines designed as follows:

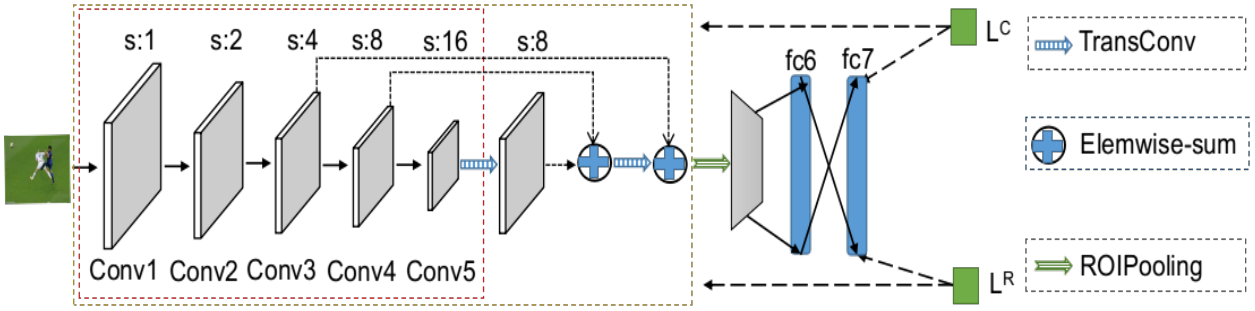
- SSD: a single stage detector [28]
- Faster R-CNN: a two-stage detector [57]
- Faster R-CNN (F_8): which directly used the feature map with a stride of eight to detect faces
- Faster R-CNN (T_{16}): which was obtained by applying *Trans-Conv* to enlarge F_{16} with double magnification as F_8
- Faster R-CNN (E_8): which was obtained by applying *Element-Wise-Sum* operation (*Sum*) to fuse the features of T_{16} and feature map with F_8, and
- Faster R-CNN (E_4), which was obtained by applying *Element-Wise-Sum* operation to fuse the features of T_8 and F_4.



(a)



(b)



(c)

Figure 3.10: Depiction of three baselines for investigating the effect of feature maps of different resolutions and their combination on the performance of detecting multi-scale faces: (a) Faster R-CNN (T_{16}), (b) Faster R-CNN (E_8), and (c) Faster R-CNN (E_4). Each baseline is based on Faster R-CNN, which is a two-stage detector. The first stage is used to propose face bounding boxes, while the second stage refines the results of the first stage. Therefore, each stage includes one classification loss and one regression loss. The symbol Element-wise-sum denotes the Element-wise-sum operation. They symbol S:N indicates that the stride of the feature map is N. The symbols L^C and L^R denote the classification loss and the regression loss, respectively.

Analysis: In our comparative study, all methods used a single feature map to predict face locations. Based on the experimental results depicted in Table. 3.9, the following observations held:

Table 3.9: Comparative study of baselines on the easy, medium and hard subsets of WIDER FACE validation set in terms of mean average precision (mAP - %).

Methods	Easy	Medium	Hard
SSD-face [28]	89.9%	85.4%	62.5%
Faster R-CNN-face [57]	89.5%	87.1%	71.6%
Faster R-CNN(F_8)	89.9%	88.6%	75.9%
Faster R-CNN(T_{16})	90.0%	88.8%	**76.4**%
Faster R-CNN(E_8)	**90.2**%	**89.0**%	74.5%
Faster R-CNN(E_4)	89.2%	88.1%	73.6%

1. By comparing the first two rows in Table 3.9, we could observe that Faster R-CNN achieved similar performance to SSD on the easy subset but far better performance on the medium and hard subsets. This illustrates that a two-stage face detector has great potential in tackling small face detection tasks.

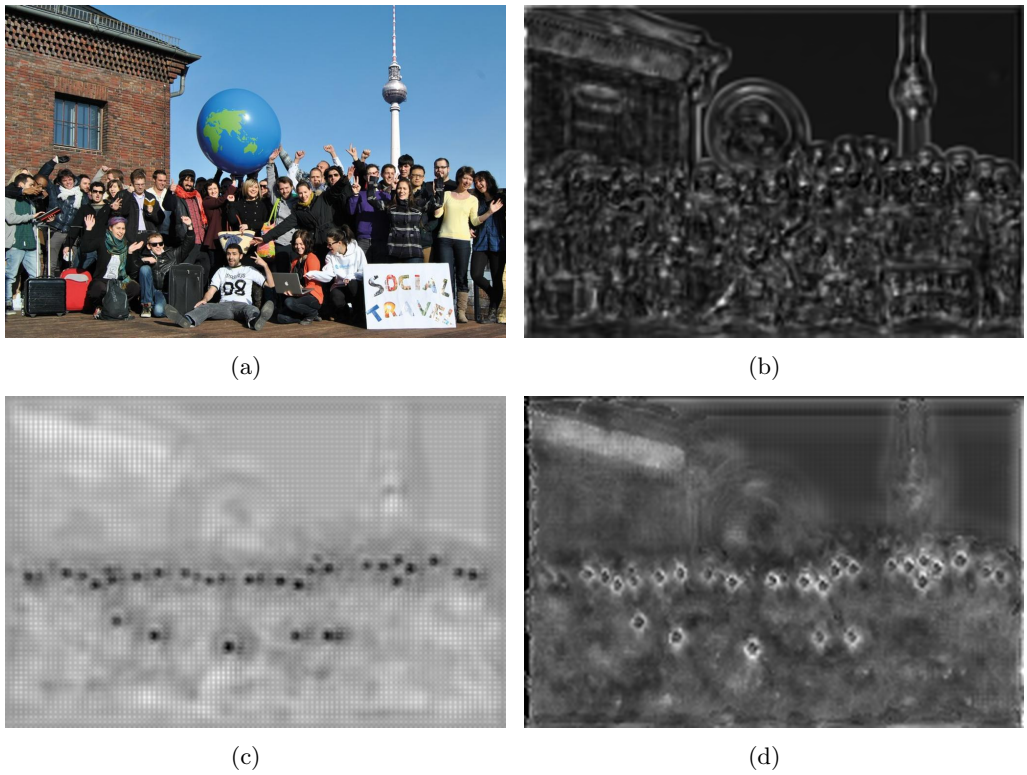
2. The performance of Faster R-CNN (T_{16}) surpassed that of Faster R-CNN (F_8) in all subsets, which indicated that networks can learn more discriminative features in deep layers than shallow layers.

3. Faster R-CNN (E_8) produced worse results than Faster R-CNN (T_{16}) in the hard set, but better in the easy and medium subsets. We posited that *Elem-Wise-Sum* is not a good way to fuse the features for detecting small faces. The above conclusions are illustrated in Fig.. 3.11, where faces have larger contrast with the background in the feature map of the deep layer, but faces and background have smaller differences in the feature maps of the shallow layer and the fused layer for small faces.

4. Faster R-CNN (E_4) exhibited overall worse performance than Faster R-CNN (E_8), which illustrated that under the feature fusion of *Elem-Wise-Sum*, the shallow layer ($conv3$) in the

backbone deteriorated the face detection performance. Therefore, the shallow layer is not always contributing to detecting multi-scale faces, especially small faces.

Fig. 3.11 depicts the feature maps of designed models in the comparative study, which also demonstrated our analysis.



(a) (b)

(c) (d)

Figure 3.11: Depiction of feature maps from different layers and their combinations. Each feature map was obtained by computing average values in the channel direction. (a) Sample image from the WIDER FACE dataset. (b) Depiction of the feature map generated by Faster R-CNN (F_8). This feature map originates from the shallow layer and includes most of the facial details, but it also contains much more noise due to indistinct edges between the faces and the background. (c) Depiction of the feature map generated by Faster R-CNN (E_8). This feature map was produced by a combination of shallow and deep layers. Although it depicts a clear edge between faces and the background, the difference between occluded as well as small faces and background is still insignificant. (d) Depiction of the feature map generated by Faster R-CNN (T_{16}). This feature map was obtained by the deep layer and exhibits the least non-face noise because the difference of features between faces and the background is distinct. Upon review of the feature maps, it appears that the feature map depicted in (d) can be used to distinguish faces of all scales better than the other two feature maps.

3.2.4.2 Evaluation on the Benchmarks

AFW Dataset [73]: Fig. 3.12(a) depicts the precision-recall curves of SSFD$^+$ on this dataset. One could observe that SSFD$^+$ achieved average precision of 99.53% and outperformed all compared methods by a large margin. Fig. 3.14(a) depicts the selected examples of SSFD$^+$ on AFW.

PASCAL Faces Dataset [54]: Fig. 3.12(b) depicts precision-recall curves of SSFD$^+$ and other compared methods on this dataset. The precision-recall curves indicated that SSFD$^+$ achieved the best performance among other compared face detectors and outperformed average precision of 6.13% at least and the best average precision of 98.24%. Fig. 3.14(b) depicts selected results of SSFD$^+$ on PASCAL Faces.

FDDB Dataset [17]: Fig.s 3.12(c,d) depict the scores of SSFD$^+$ on FDDB dataset. FDDB faces were annotated by ellipse coordinates, while SSFD$^+$ produced rectangular face coordinates, which has a great impact on the continuous test. However, SSFD$^+$ achieved discontinuous and continuous scores of 97.5% and 75.6%, respectively and outperformed HR, Faster R-CNN and ScaleFace on both discontinuous and continuous evaluations. In additional, SSFD$^+$ had a small gap with SFD on discontinuous evaluation.

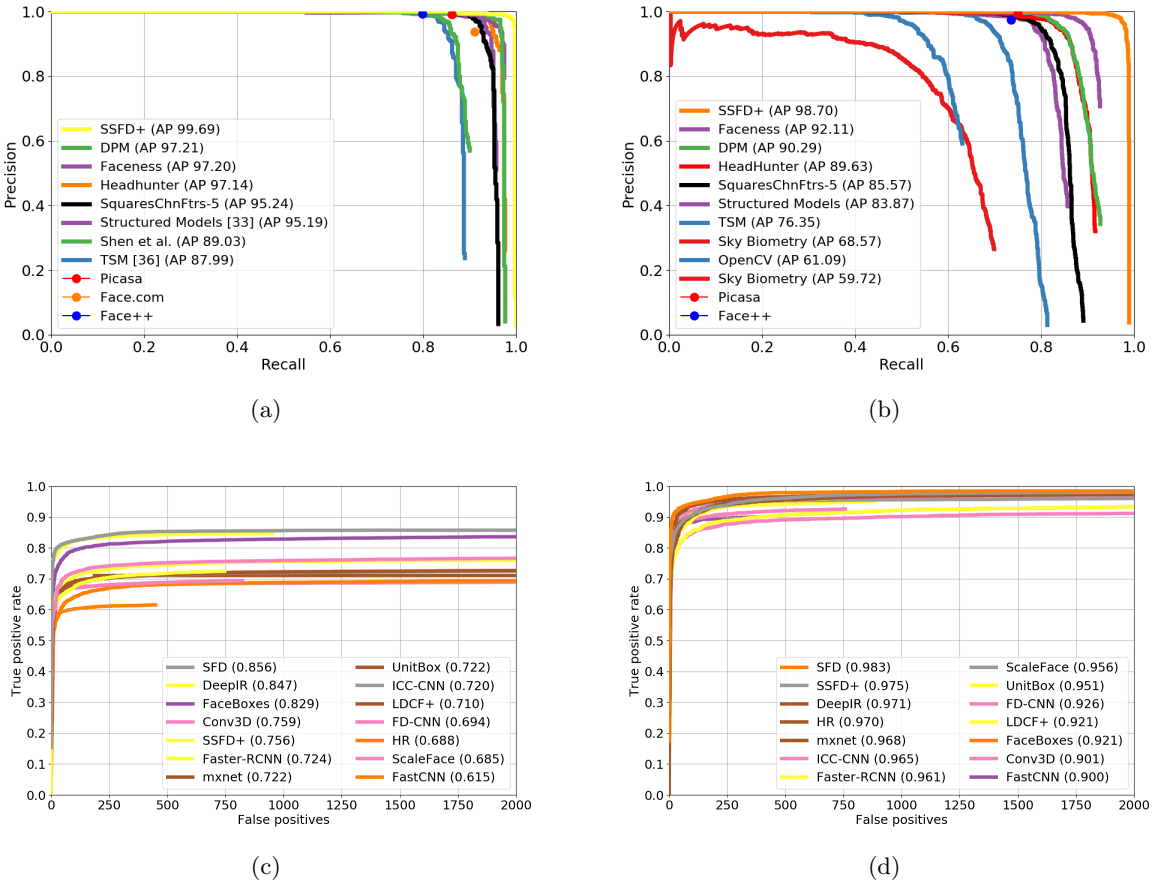


Figure 3.12: Evaluation results of SSFD$^+$ on the AFW, PASCAL faces, and FDDB datasets. (a) precision-recall curves on AFW; (b) precision-recall curves on PASCAL faces; (c) continuous score curves on FDDB, and (d) discontinuous score curves on FDDB.

WIDER FACE Dataset [55]: Fig. 3.13 depicts the precision-recall curves and mAP results of SSFD$^+$ for the WIDER FACE validation and test sets.

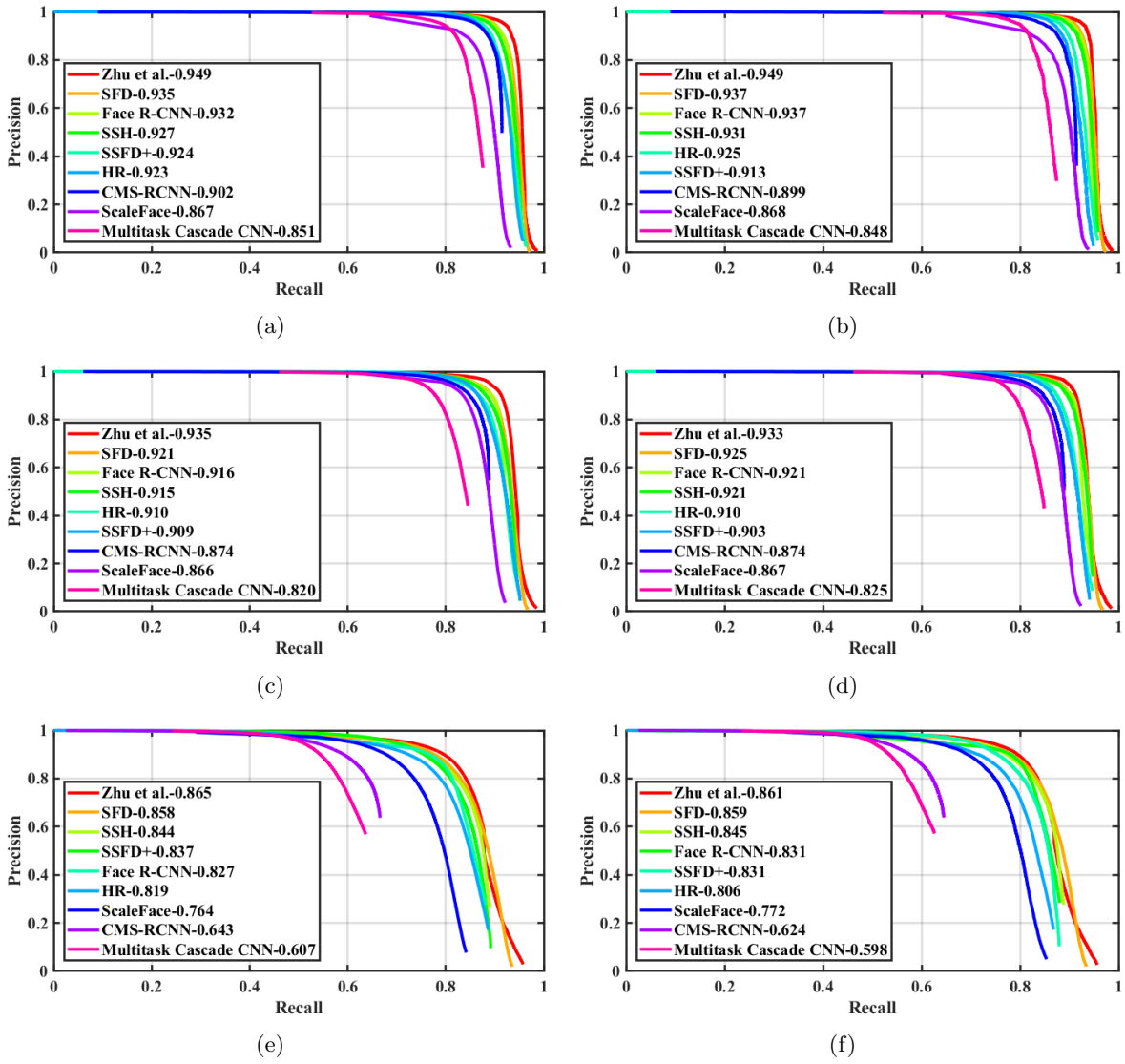


Figure 3.13: Precision-recall curves of SSFD$^+$ on the WIDER FACE testing (a, c, e) and validation (b, d, f) sets: (a,b) Easy, (c,d) Medium, and (e,f) Hard, respectively.

SSFD$^+$ achieved mAPs of (91.3%, 90.3%, 83.1%) and (92.4%, 90.9%, 83.7%) on the (easy, medium and hard) subsets of WIDER FACE validation and testing sets, respectively. It's worth mentioning that SSFD$^+$ achieved the same or better performance than Face R-CNN on the hard subset of the WIDER FACE validation and testing sets. Face R-CNN employed both multi-scale training and multi-scale testing. However, SSFD$^+$ deployed a multi-scale testing strategy to improve the performance. Therefore, SSFD$^+$ was an effective method of detecting small faces. In addition, SSFD$^+$ with VGG16 backbone network outperformed VGG16-HR with input pyramid, and ScaleFace with ResNet-101 backbone network on all subsets of WIDER FACE validation and testing set, which demonstrated the effectiveness of SSFD$^+$ on detecting multi-scale faces. Moreover, mAPs of SSFD$^+$ on all subsets of WIDER FACE testing set were better than that of WIDER FACE validation set, which indicated the robustness of SSFD$^+$ on localizing multi-scale faces, especially small faces.

Fig. 3.14(d) depicts selected qualitative results of SSFD$^+$ on the WIDER FACE validation and test sets, which cover blurry, large, small and occluded faces with various poses. We could observe that SSFD$^+$ could accurately localize faces with a variety of scales in extreme environments. Finally, selected failure cases are depicted in Fig. 3.15. The failures mainly resulted from the extreme pose, extreme occlusion, illumination and small size of faces. The hardest case was that the smallest size faces we can't detect. It was difficult for the network to extract enough discriminaticve facial features from small faces. Although facial information can be increased by magnifying small face regions, it also introduced more noise.

(a) (b)

(c) (d)

Figure 3.14: Qualitative results of SSFD$^+$ on benchmarks. (a) AFW; (b) PASCAL faces; (c) FDDB; (d) WIDER FACE.

Figure 3.15: Failure cases of SSFD$^+$. The bounding boxes in green and in red indicate ground truth and predicted faces, respectively.

Table 3.10: SSFD$^+$ inference time compared to other methods.

Method	SSH [33]	Faster R-CNN [38]	SSFD$^+$	HR [16]
Time (ms)	182	335	582	1,010

Inference Time: Table 3.10 depicts the inference time of SSFD$^+$ and compared methods. All methods shared the same input setting. SSFD$^+$ was timed on a single Tesla V100 GPU. The shorter side of the input image was $1,200$ pixels, the longer side of the input image was $1,600$ pixels. The average inference time was computed for the whole of the WIDER FACE validation set. We noted that SSFD$^+$ outperformed HR also in inference time (582 ms/image vs. 1010ms/image). It is worth of mentioning that SSFD$^+$ outperformed HR in terms of performance and inference time. As expected, Faster R-CNN performed inference in shorter time then SSFD$^+$. This is because SSFD$^+$ was based on Faster R-CNN and the addition of a transposed convolutional layer resulted in slower inference speeds due to the increased resolution of feature maps. Finally, in comparison to SSH, SSFD$^+$ achieved improved performance (in terms of mAP) on the hard subset of the WIDER FACE validation set, at the cost of increased inference time.

3.3 SANet: Smoothed Attention Network for Single-stage Face Detector

3.3.1 General Architecture

Fig. 3.16 depicts the architecture of SANet, which inherited the scale-equitable framework of S^3FD [70]. Moreover, the max-in-out of PyramidBox [44] was applied to each detection layer and image

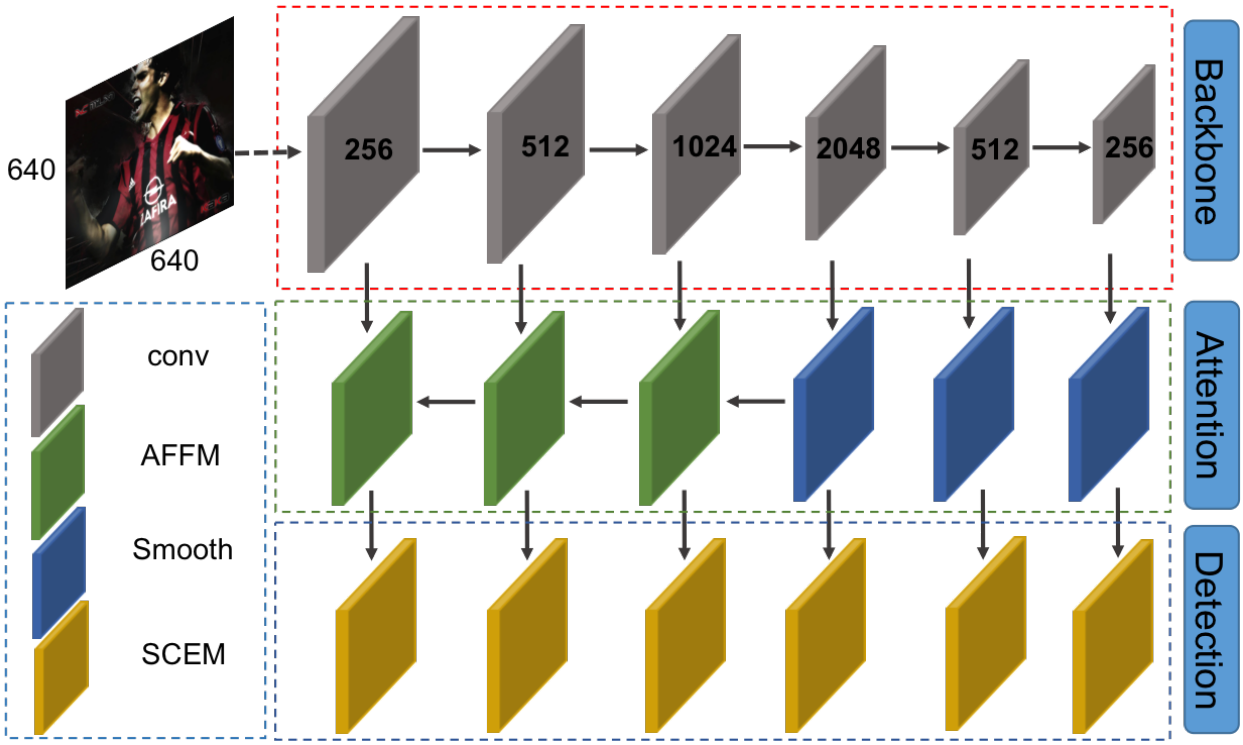


Figure 3.16: Depiction of the architecture of SANet. AFFM was used to effectively integrate low-level and high-level semantic features by applying attention mechanism to high-level semantic features to decrease noise in the fused features. SCEM was leveraged to incorporate additional contextual information by concatenating features from different receptive fields and circumvent the gridding artifacts problem caused by dilated convolution.

cropping was used for data augmentation. The modified S^3FD network was denoted as S^3FD-M. The backbone layers consist of feature maps with strides of 4, 8, 16, 32, 64 and 128, respectively. The AFFMs and smoothed layers were applied to backbone layers to obtain fused features and smoothed features. The fused features and smoothed features were then individually fed into the SCEM to obtain detection layers, which were deployed to predict face bounding boxes and confidence scores.

3.3.1.1 Attention-guided Feature Fusion Module

Fig. 3.17 depicts the architectures of AFFM and its four variants. To keep the channel consistency

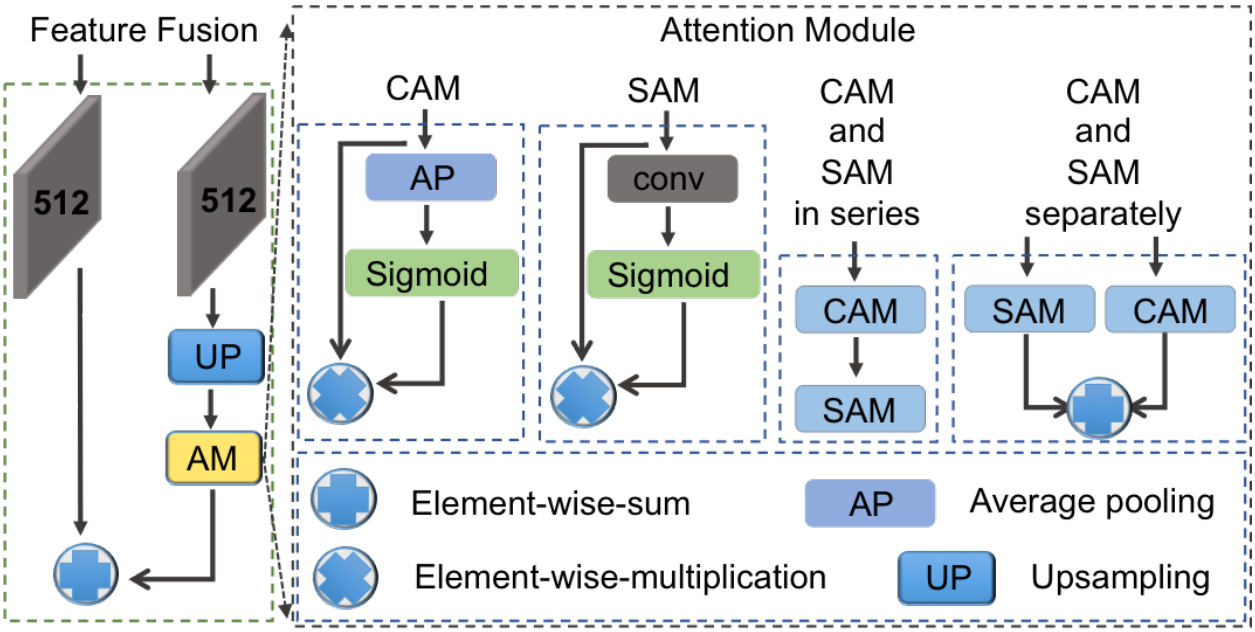


Figure 3.17: Depiction of the architecture of AFFM. Two convolution layers converted the initial feature maps to 512 channels to improve computational efficiency. Upsampling was then used to enlarge the feature map at the high-level to be the same size as the feature map at the low-level. To obtain attention-focused features, the AFFM applied the attention module to the magnified high-level feature map. Finally, to fuse the attention-focused features and the low-level features, element-wise-sum was implemented. Four variants of the attention module take channel-wise attention, spatial-wise attention, series, and distinct combinations into consideration.

of input feature maps, two convolutional layers with 512 channels were applied to the high-level and low-level feature maps, respectively. To keep the size consistency of input feature maps, the upsampling operation was used to magnify the high-level feature map to the same size as the low-level feature map. Then the magnified high-level feature map was fed into the attention module to obtain the attention-focused feature map, where were merged into the low-level feature map finally. Because the attention module ensures the network focuses on reliable information and disregards noise, it can decrease noise in the fused features. We also intended four AFFM variants

to investigate the impact of the attention mechanism on performance, taking into account channel-wise attention, spatial-wise attention, and their combinations.

Channel-wise Attention Module: Let's assume that the input feature map has the $C \times H \times W$ dimension as C feature maps with $H \times W$, where C, H, and W denote the number of channels, height, and width respectively. Global average pooling ($G(\cdot)$) was applied channel-wisely to each feature map to improve the representation power of inter-channel features. A feature map with $C \times 1 \times 1$ was then obtained, which was followed by a sigmoid function ($S(\cdot)$) to get C attention weights corresponding C channels of the input feature map. To get a channel-wide attention-focused feature map (F_c), multiplication between C attention weights and the input feature map was applied. The process can be formulated as

$$F_c = S(G(F)) * F, \tag{3.2}$$

Spatial-wise Attention Module: To increase representation power of inter-spatial features, one convolutional layer (C_v) was applied to the input feature map with $C \times H \times W$ to obtain a feature map with $1 \times H \times W$, which was followed by a sigmoid function ($S(\cdot)$) to get spatial attention weights. Multiplication was then applied to the input feature map and spatial-wise attention weights to get a spatial-wise attention-focused feature map (F_s). This process can be formulated as follows:

$$F_s = S(C_v(F)) * F, \tag{3.3}$$

where C_v denotes the convolutional layer with kernel size of 3×3, padding of 1, stride of 1 and output channel of 1.

Attention Feature Fusion: Channel-wise and spatial-wise attention-focused features represent features in different spaces. To explore the effect of attention-focused feature fusion on performance, we designed two variants to combine channel-wise and spatial-wise attention-focused features. The first variant applied the sequence of channel-wise and spatial-wise attention modules. The output of the channel-wise attention module was directly used as the input of the spatial-wise attention

module. This process is denoted as follows:

$$F_m = S(C_v(F_c)) * F_c, \tag{3.4}$$

where F_m represents fused attention-focused feature map with $C \times H \times W$.

The second variant as to apply the channel-wise and spatial-wise attention modules to the low-level and high-level semantic features, respectively. Specifically, the spatial-wise attention module was applied to the low-level semantic features to obtain the spatial-wise attention-focused features, The channel-wise attention module was applied to the high-level semantic features to obtain the channel-wise attention-focused features. Then the element-wise-sum was applied to the channel-wise attention-focused features and the spatial-wise attention-focused features to get the fused features (F_e). This process can be computed as follows:

$$F_e = S(C_v(F_l)) * F + S(G(F_h)) * F, \tag{3.5}$$

where F_l is the low-level semantic feature map, F_h is the high-level semantic feature map.

3.3.1.2 Smoothed Context Enhancement Module

Fig. 3.18 depicts the design of the SCEM. The SCEM was designed to enrich contextual information and solve the gridding artifacts problem.

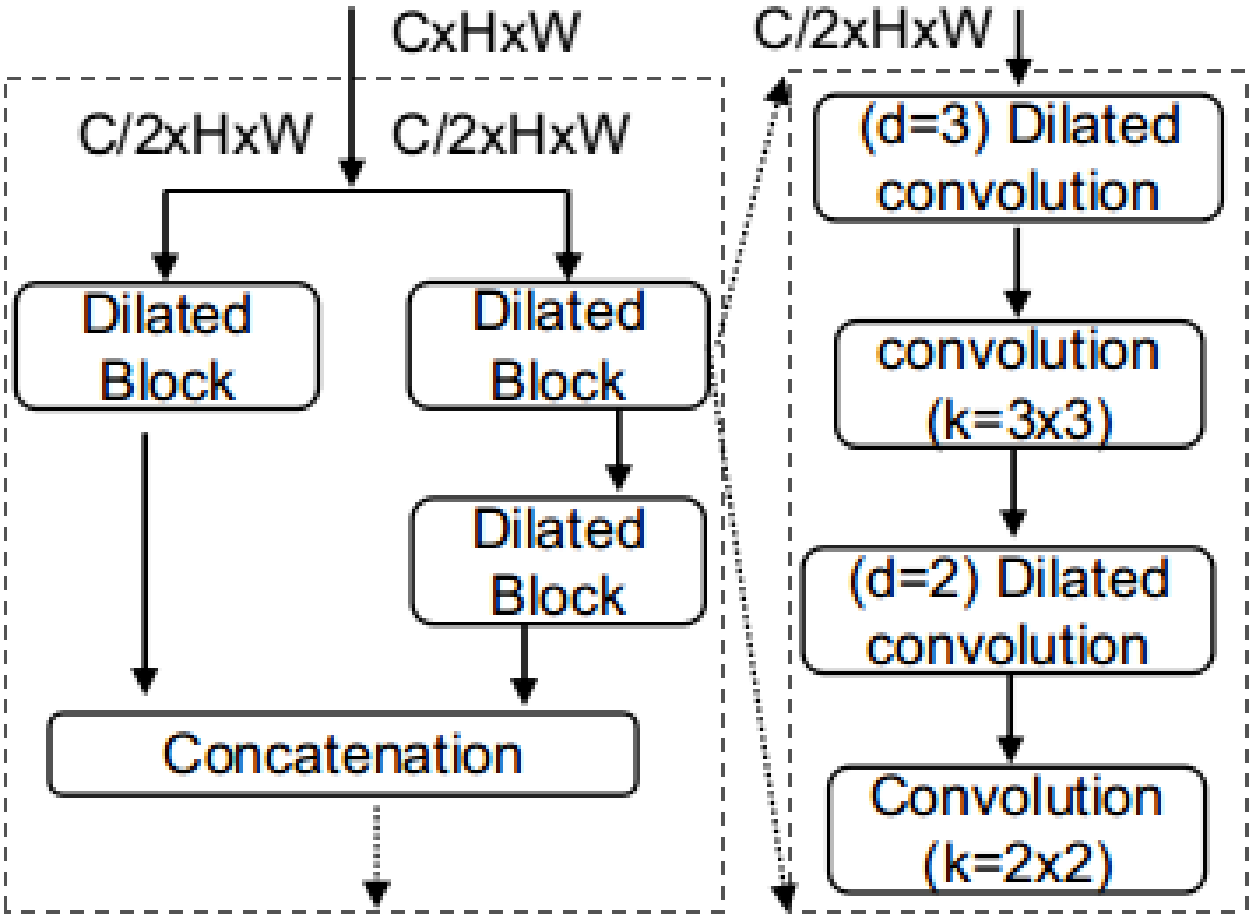


Figure 3.18: Depiction of the architecture of SCEM. In the Dilated Block, these two dilated convolutions have the kernel size of 3×3 and 2×2, the dilation rate of 3 and 2, the padding of 3 and 2, respectively, and the same stride of 1. These two convolutional layers have the same stride of 1, padding of 1 and different kernel sizes of 3×3 and 2×2, respectively.

Compared to the classic convolution with the same kernel size, the dilated convolution magnified the receptive field by interlacing 0s with the number of $r - 1$ in the middle of adjacent kernel elements, where r was the dilation rate. For instance, a dilated convolution has a kernel size of 3×3 and a dilation rate of 2. Because of the interlacing of one 0 in the middle of adjacent kernel elements,

if the stride parameter is equal to 1, four adjacent elements in the dilated convolution feature map could be obtained from completely different unit sets in the input feature map. Therefore, the gridding artifacts problem result in inconsistency of local information. Fig. 3.19 depicts an example of the gridding artifacts problem [51]. To address the issue of gridding artifacts, in each

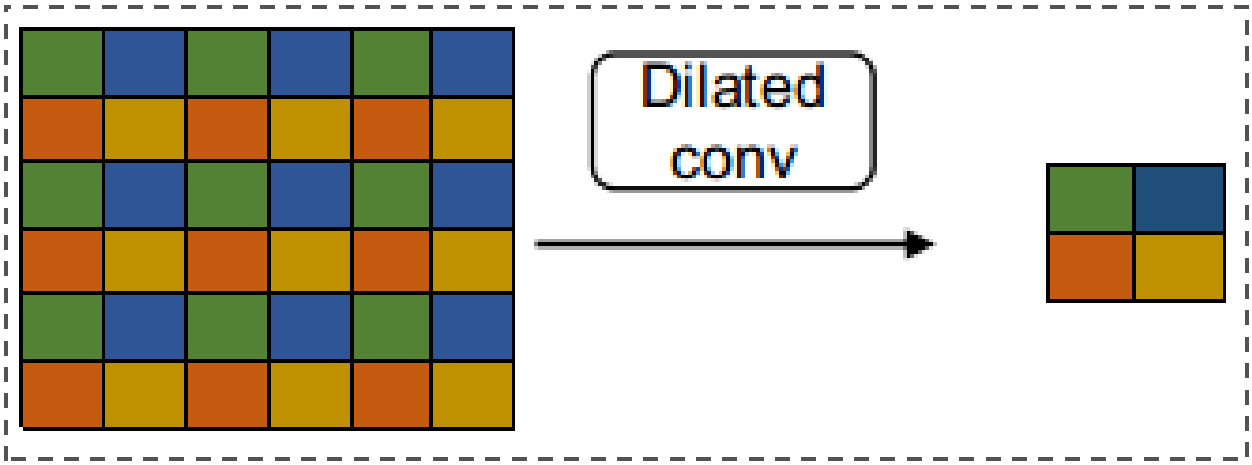


Figure 3.19: Demonstration of issue of gridding artifacts. Dilated conv denotes the dilated convolution. The dilated convolution has a kernel size of 3×3, stride of 1, and dilation rate of 2. The green pixels in the feature map on the right side are acquired by nine green pixels in the feature map on the left side. The pixels act as above with other distinct colors. Neighboring four pixels in the feature map on the right side are therefore acquired in the feature map on the left side by totally different four sets of units.

Dilated Block, one dilated convolution was followed by one convolutional layer to re-learn the local relationship between completely separate sets of units obtained by the dilated convolution. The kernel size of the convolutional layer was equal to the number of the neighboring irrelevant elements in the feature map produced by the dilated convolution. Dilated convolutions with dilation rates of 3and 2×2 were followed in specific by convolutional layers with kernel sizes of 3×3 and 2×2, respectively. Finally, to enrich contextual information, we stacked Dilated Blocks and concatenate features from different receptive field size. Compared to other methods using dilated convolution in context module, our module can effectively retain local information consistency and elevate computational efficiency of the magnifying receptive field to enrich contextual information.

3.3.2 Implementation Details

Training and Inference: The weights of ResNet-50 pre-trained model on the ImageNet [39] is used to initialize the weights of the backbone network. Other layers are initialized by using the xavier initializer. The scales of tiled anchors are set to [16, 32, 64, 128, 256, 512]$\times\sqrt[3]{2}$ with the ratio of 1. In the training stage, the IoU threshold value used to assign anchors to be positive and negative samples is 0.35, and the online hard example mining technology is applied to all experiments. The ratio of positive and negative samples is 1:3. Batch size is set to 8 and 14 with normalized inputs of 640×640 in the ablation study and evaluation on benchmarks experiments, respectively. SGD with the momentum of 0.9, gamma of 0.1 and weight decay of 5×10^{-4} are used to train our model. In total, there are 120K iterations, and the learning rate is set to 0.001 initially. After 80K and 100K iterations, the learning rate is multiplied by 0.1. During the inference, the threshold value of nms is set to 0.5. We follow S^3FD [70] to use multi-scale testing strategy to evaluate our methods.

Loss Function: A multi-task loss function is defined in terms of classification and regression tasks. In particular, the loss function is computed as follows:

$$L\left(p_i, r_j\right) = \frac{1}{M}\left(\sum_{i}^{M} L_c\left(p_i, p_i^*\right) + \sum_{j}^{M} p_i^* L_r\left(r_j, r_j^*\right)\right) \tag{3.6}$$

where L_c is the softmax cross-entropy loss over face and background classes, L_r is the smoothed L_1 loss for the bounding boxes regression task. In the first term of the sum, p_i is the probability predicted for an anchor that belongs to the face or background classes, p_i^* is the ground truth label for the classification task, if an anchor belongs to face class, the p_i^* is set to 1, otherwise to 0. In the second item of the sum, the symbol r_i denotes the predicted parametric coordinates following Girshick *et al.* [12] and r_i^* is the corresponding ground truth. Finally, M is the mini-batch size.

3.3.3 Evaluation Databases and Metrics

The precision-recall curves and mean average precision (mAP) were used to evaluate all models.

AFW Dataset: It contains 473 labeled faces with large variations of pose and scale for 205 images. SANet was evaluated against methods [56] [32] [54] and commericial face detectors (e.g. Face.com, Face++ and Picasa).

PASCAL Face Dataset: It consists of 851 images and $1,335$ annotated faces with large pose and face appearance variations. As a subset, it was selected from PASCAL dataset with person layout.

FDDB Dataset: It has $5,171$ labeled faces for $2,845$ images. We evaluated SANet against to state-of-the-art models [19] [44] [70] [62] [50] [38] [56] [16] [57] [64] [46] [25] on the FDDB dataset.

WIDER FACE Dataset: The WIDER FACE training, validation and testing sets include $12,880$, $3,226$, $16,290$ images and $158,989$, $39,496$, $192,518$ annotated faces, respectively. Moreover, validation and testing sets were categorized to three subsets in terms of levels of difficulty: easy, medium and hard, which was based on the detection rate of EdgeBox [74].

UFDD Dataset: Practical conditions such as weather-based degradations, different types of blur and distractor images are considered in this dataset. UFDD consists of $6,424$ images with $10,895$ face annotations and it involves the following key conditions: (i) Rain, (ii) Snow, (iii) Haze, (iv) obscurants that appear between the object and camera lens, (v) Blur, (vi) Illumination variations, and (vii) Distractors.

3.3.4 Experimental Results

3.3.4.1 Evaluation on the Benchmarks:

AFW Dataset: Fig. 3.20(a) depicts the precision-recall curves of the SANet and compared methods on the AFW dataset. SANet achieved mAP of 99.12% and outperformed all compared methods by at least mAP of 1.91%.

PASCAL Face Dataset: As illustrated in Fig. 3.20(b), SANet outperformed all compared methods [56] [32] [54] and some commercial and library face detectors (e.g. Sky Biometry, Face++ and Picasa, OpenCV) with mAP of 99.22%.

FDDB Dataset: Fig. 3.20(c,d) depict the evaluation results of SANet and other methods on the FDDB dataset. SANet achieved 98.6% and 76.4% for discontinuous and continuous scores, respectively. SANet outperformed SFD in discontinuous scores, while performed worse than SFD in continuous scores. However, SANet produced rectangular face coordinates when SFD trained an elliptical regressor to transform their predicted bounding boxes to bounding ellopses, which has an impact on final results.

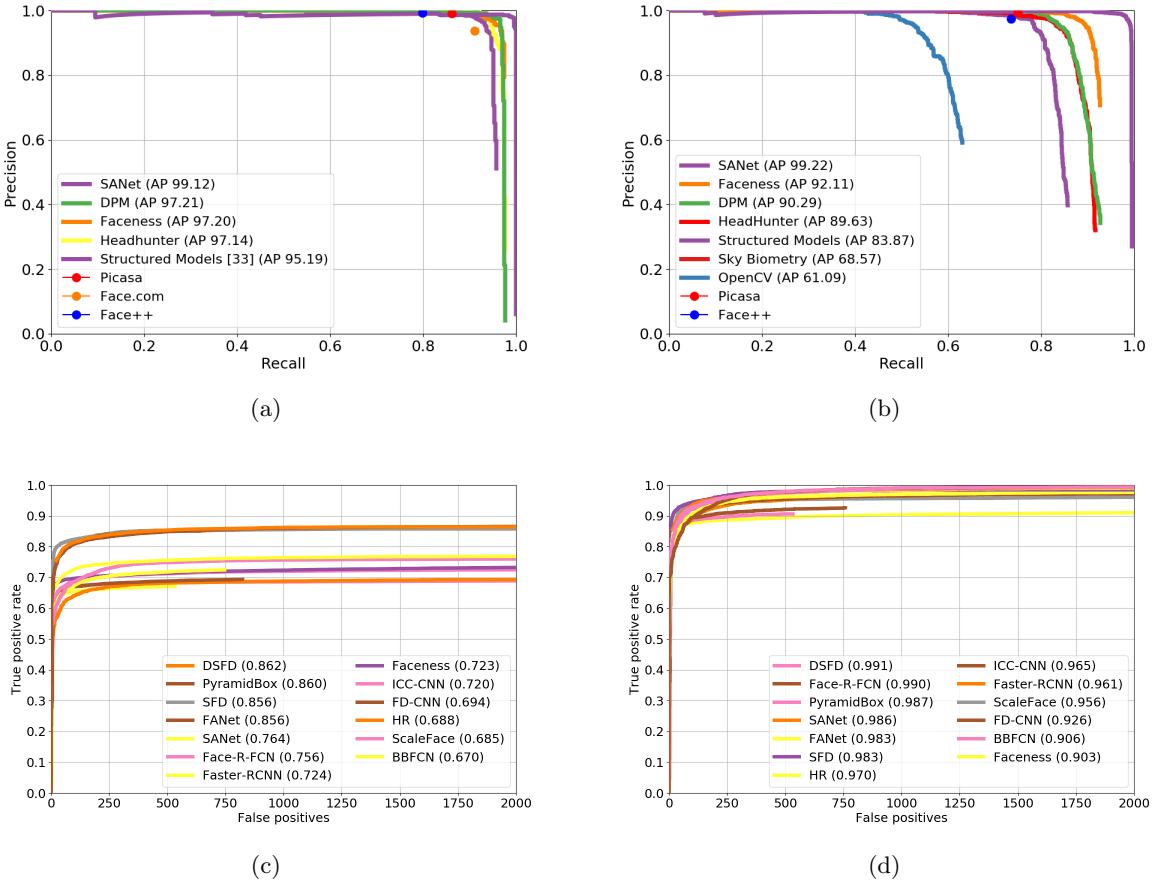


Figure 3.20: Evaluation results of the SANet on the AFW, PASCAL Faces, and FDDB datasets. (a) precision-recall curves on the AFW dataset; (b) precision-recall curves on PASCAL Face dataset; (c) continuous score curves on the FDDB dataset, and (d) discontinuous score curves on the FDDB dataset.

WIDER FACE Dataset: Fig. 3.21 depict the precision-recall curves of SANet on the WIDER FACE validation and testing sets against to the state-of-the-art face detectors [22] [45] [62] [70] [16] [57] [33] [47] [68]. The SANet achieves mAPs of 94.6%, 93.8%, 88.2% on the easy, medium and hard

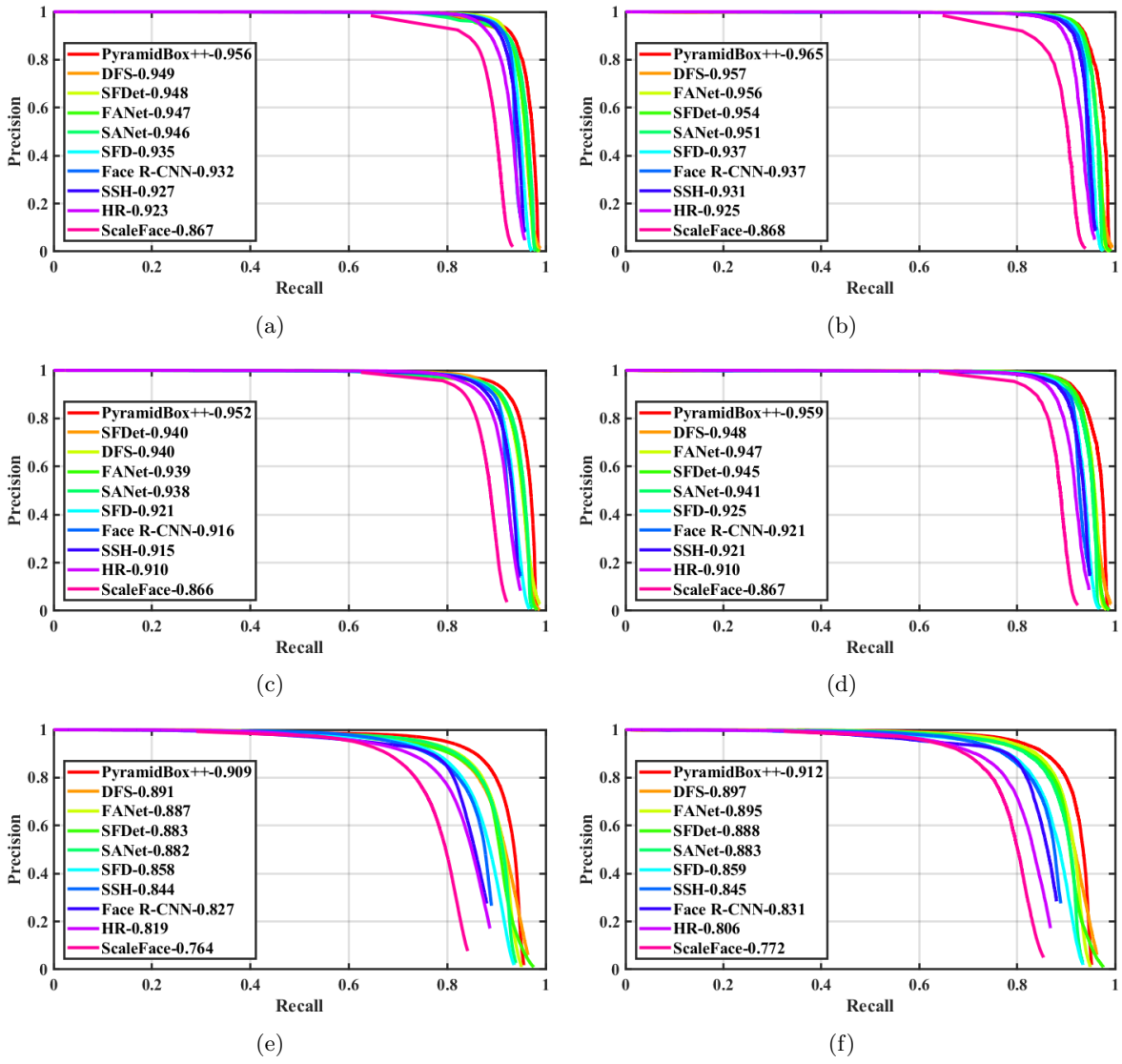


Figure 3.21: Precision-recall curves of SANet on the WIDER FACE testing (a, c, e) and validation (b, d, f) sets: (a,b) Easy, (c,d) Medium, and (e,f) Hard, respectively.

subsets respectively. It's worth noting that the SANet outperformed S^3FD and SSH on all subsets with at least mAP of 2.4%, SANet outperformed Face R-FCN and Face R-CNN on all subsets,

which were typically two one-stage and two two-stage face detectors. In addition, the difference of performances of SANet on the hard subsets of WIDER FACE validation and testing sets was only 0.1%, which demonstrated the robustness of the SANet on detecting small faces.

UFDD Dataset: Fig. 3.22 depicts the precision-recall curves of SANet on the whole of UFDD dataset with the distractor against to [70] [16] [33] [38]. We could observe that SANet outperformed

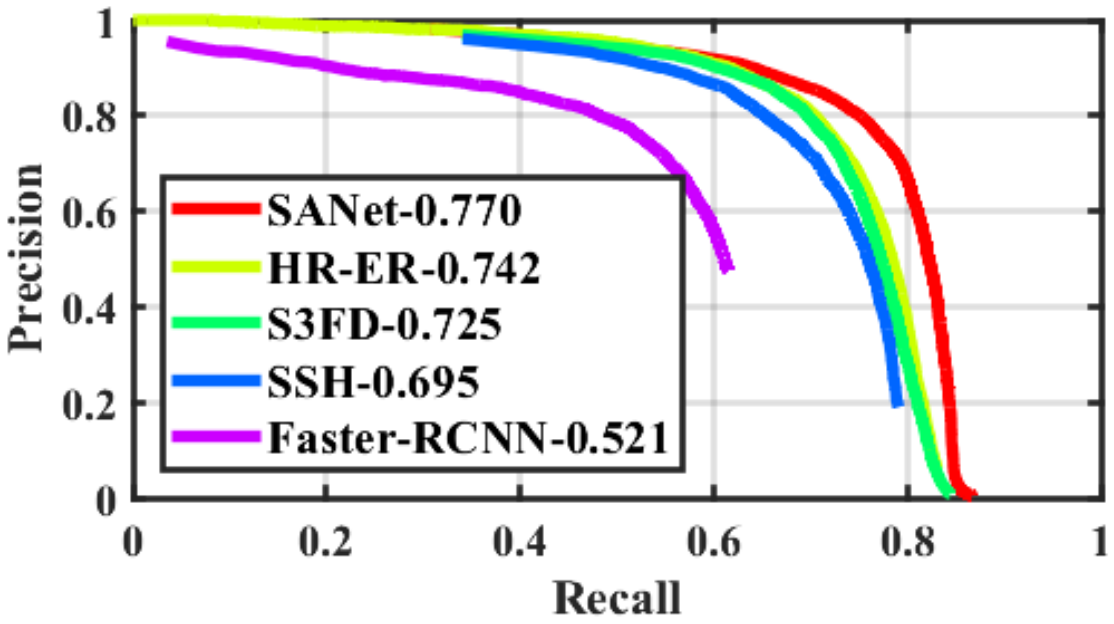


Figure 3.22: The precision-recall curves of the SANet on the UFDD dataset.

all baselines in this dataset and achieves 0.77 of mAP, which fully demonstrated the robustness and superiority of SANet on detecting multi-scale faces in a complex environment. Table 3.11 depicts the performance of the SANet on each condition of the UFDD dataset. It is worth noting that

Table 3.11: mAP% results of SANet on each condition of the UFDD dataset.

Methods	Rain	Snow	Haze	Blur	Illumination	Lens impediments	without-distractor
Faster R-CNN [38]	54.8	54.9	46.4	68.0	57.9	52.6	56.4
SSH [33]	73.5	71.3	65.4	80.6	72.0	59.4	72.5
S3FD [70]	75.9	72.3	71.9	83.8	78.0	60.7	76.1
HR-ER [16]	75.9	74.3	72.5	84.4	77.2	68.5	76.7
SANet	**78.7**	**77.2**	**75.3**	**87.8**	**82.7**	**69.4**	**80.2**

SANet achieved the best performance than all compared approaches on each condition of the UFDD

dataset. Specifically, SANet achieved mAPs of 78.7%, 77.2%, 75.3%, 87.8%, 82.7%, 69.4%, 80.2% on the Rain, Snow, Haze, Blur, Illumination, Lens impediments, and without-distract conditions, respectively. The results well proved the strong generalization ability of the SANet on detecting faces with various scales in a complex environment.

Inference Time: Table 3.12 depicts the inference time of SANet and compared methods. SANet

Table 3.12: SANet inference time compared to other methods.

Method	SSH [33]	SANet	SSFD$^+$	HR [16]
Time (ms)	182	570	582	1,010

was timed on a single Tesla V100 GPU. A single-scale input image was used as input of all compared methods. The shorter side of the input image was $1,200$ pixels, the longer side of the input image was $1,600$ pixels. The average inference time was computed for $3,226$ images of the WIDER FACE validation set. It is worth of mentioning that SANet outperformed the two-stage face detector, SSFD$^+$ designed by us, in inference time (570 ms/image vs. 582 ms/image). Although SANet employed multi-scale feature maps and SSFD$^+$ used a single-scale feature map, SANet performed inference in a shorter time and better than SSFD$^+$ on detecting multi-scale faces. Finally, SANet performed better than SSH in terms of mAP on the hard subset of the WIDER FACE validation set at the cost of increased inference time.

3.3.4.2 Comparative Study

On one hand, the impact of four variants of the attention module on the performance was examined, which include channel-wise, spatial-wise attentions and their combinations. On the other hand, the influence and role of the SCEM on improving the performance were investigated. All comparative were trained on the WIDER FACE training set and evaluated on the WIDER FACE validation set. All extensive experiments shared the same parameters setting and were conducted on a single GPU.

Comparative Study: AFFM To investigate the impact of the attention mechanism on the performance, we evaluated four variants of the AFFM and baseline and analyzed their results:

1. S^3FD-M: our baseline;
2. S^3FD-M+CA: AFFM only includes channel-wise attention module;
3. S^3FD-M+SA: AFFM only includes spatial-wise attention module;
4. S^3FD-M+(CA-SA): channel-wise attention module and spatial-wise attention module were applied in series in the AFFM;
5. S^3FD-M+(CA+SA): channel-wise attention module and spatial-wise attention module were applied separately in the AFFM.

Analysis: Table 3.13 summarizes the results of the comparative study of the AFFM. One could

Table 3.13: mAP% results of the comparative study of the AFFM on the WIDER FACE validation set.

Methods	Validation set		
	Easy	Medium	Hard
S^3FD-M	93.7	91.8	81.0
S^3FD-M+CA	94.7	93.5	86.4
S^3FD-M+SA	94.8	93.4	86.8
S^3FD-M+(CA-SA)	94.3	93.3	86.1
S^3FD-M+(CA+SA)	94.4	93.2	86.3

observe that the performance of all designed models outperformed the baseline mAP of 5.1% at least on the hard subset and was better than the performance of our baseline on all subsets. The results indicated that utilizing an attention mechanism is an effective way of fusing low-level and high-level semantic features and reducing noise in the fused feature map for improving the performance of detecting multi-scale faces. Moreover, compared to series and separate combinations of the spatial-wise attention and the channel-wise attention with mAPs of 94.3%, 93.3%, 86.1% and 94.4%, 93.2%, 86.3% on the easy, medium, and hard subsets respectively, applying the channel-wise attention alone and the spatial-wise attention along achieve better performance with mAPs of

94.7%, 93.5%, 86.4% and 94.8%, 93.5%, 86.8% on the easy, medium, and hard subsets respectively. This observation suggested that their combinations produce redundant information or noise. Although the channel-wise and the spatial-wise attention-focused features were learned in different feature spaces, they have many similar characteristics. Furthermore, the performance of applying the spatial-wise attention was better than the performance of applying the channel-wise attention on the hard subset, which means that the inter-spatial information is more important than the inter-channel information on detecting small faces. Intuitively, even the whole facial information is not enough for people to finding tiny faces, let alone local facial information. Therefore, the inter-spatial information includes more contextual information on small faces than the inter-channel information. We also found that applying the spatial-wise attention alone costs less memory during training and produces less bounding boxes during inference than applying the channel-wise attention alone.

Comparative Study: SCEM Compared to the attention mechanism, to explore the impact of the role of the SCEM on the performance, we design the following models and our network structure:

1. S^3FD-M+CA+SCEM: the output of the channel-wise attention module was fed into the SCEM;
2. S^3FD-M+SA+SCEM: the output of the spatial-wise attention module was fed into the SCEM;
3. S^3FD-M+CA+SCEM(SA)): the output of the channel-wise attention module was fed into the SCEM and the SCEM is replaced with the spatial-wise attention module;
4. SANet: the output of the spatial-wise attention module is fed into the SCEM, and the model was trained with a large batch size of 14.

Analysis: Table 3.14 summarizes the results of the comparative study of the SCEM. Compared to Table 3.13, one can observe that applying the SCEM to the channel-wise or the spatial-wise attention module improves the performance by 0.5%, 0.4%, 0.6% and 0.2%, 0.4%, 0.9% on the

Table 3.14: mAP% results of comparative study of the SCEM and SANet on the WIDER FACE validation set.

Methods	Validation set		
	Easy	Medium	Hard
S^3FD-M+CA+SCEM	95.2	93.9	87.0
S^3FD-M+SA+SCEM	95.0	93.8	87.7
S^3FD-M+CA+SCEM(SA)	94.3	93.0	86.4
SANet	**95.1**	**94.1**	**88.3**

easy, medium, and hard subsets respectively. These results demonstrated the effectiveness of the SCEM on improving the performance of detecting the various scale of faces. Furthermore, applying the spatial-wise attention module with the SCEM achieved better performance than applying the channel-wise attention module with the SCEM on the hard subset. To investigate the roles of the SCEM and the attention mechanism on the performance, we replaced the SCEM with the spatial-wise attention mechanism. Interestingly, it achieved the worst performance than other models. From this observation, one can obtain the conclusion: the AFFM and the SCEM play different roles and provide different facial features in detecting multi-scale faces.

Ablation Study: Gridding Artifacts To investigate the impact of the gridding artifacts problem on the performance, we adopt the following first three model:

1. SANet: the output of spatial-wise attention module is fed into the SCEM;

2. SANet(Conv)): the output of the spatial-wise attention module is fed into the SCEM; the SCEM replaces all dilated convolutions with the corresponding convolutions with the same kernel sizes to mitigate the influence of the number of parameters.

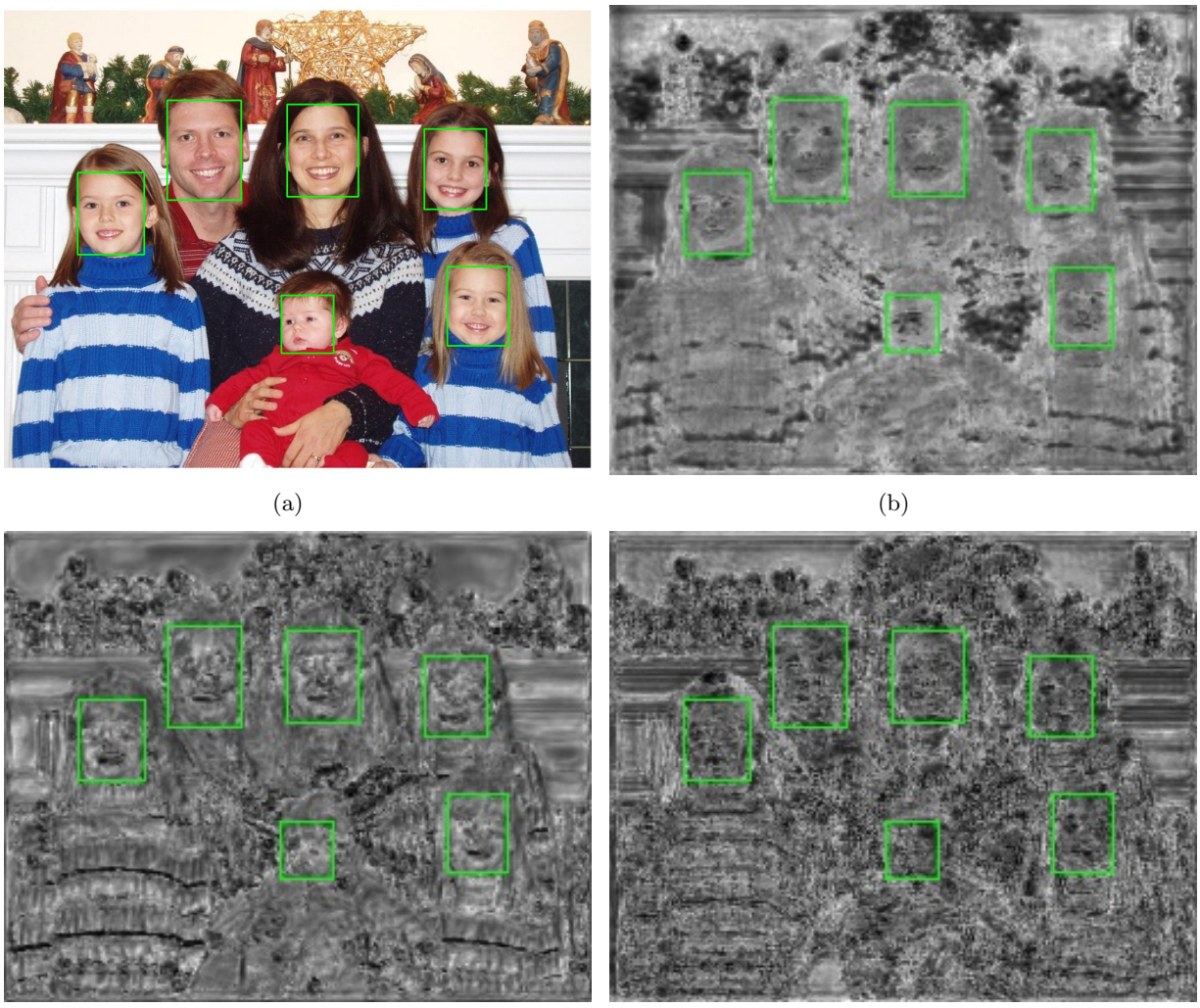
3. SANet(DConv)): the output of the spatial-wise attention module is fed into the SCEM; the SCEM replaces all convolutions with the corresponding dilated convolutions with the same kernel sizes to mitigate the influence of the number of parameters.

Analysis: Table 3.15 depicts the results. One can observe that compared to stacked dilated convolutional layers, stacked convolutional layers improve the performance with mAPs with 0.3%,

Table 3.15: mAP% results of the ablation study of the gridding artifacts problem on the WIDER FACE validation set.

Methods	Validation set		
	Easy	Medium	Hard
SANet(DConv)	90.6	88.9	80.9
SANet(Conv)	90.9	89.6	83.5
SANet	95.0	93.8	87.7

0.7% and 2.6% on the easy, medium and hard subsets, respectively, which suggests that the gridding artifacts problem has an important impact on detecting small faces. Furthermore, our approach achieves the best performance by improving the performance by mAPs of 6.7% and 4.2% on the hard subsets, respectively, which fully demonstrated the effectiveness of our approach on solving the gridding artifacts problem. The Fig. 3.23 also supports this claim. In comparison, the feature map obtained by SANet(DConv) loss a large amount of detailed facial information due to the gridding artifacts problem, which distorts the reliable facial features. Although the feature map obtained by SANet(Conv) includes more details, the contrast of the facial features and background is not discriminative. Finally, the feature map obtained by our approach includes more complete and discriminative facial features around regions such as eyes, nose, and mouth.



(a) (b)

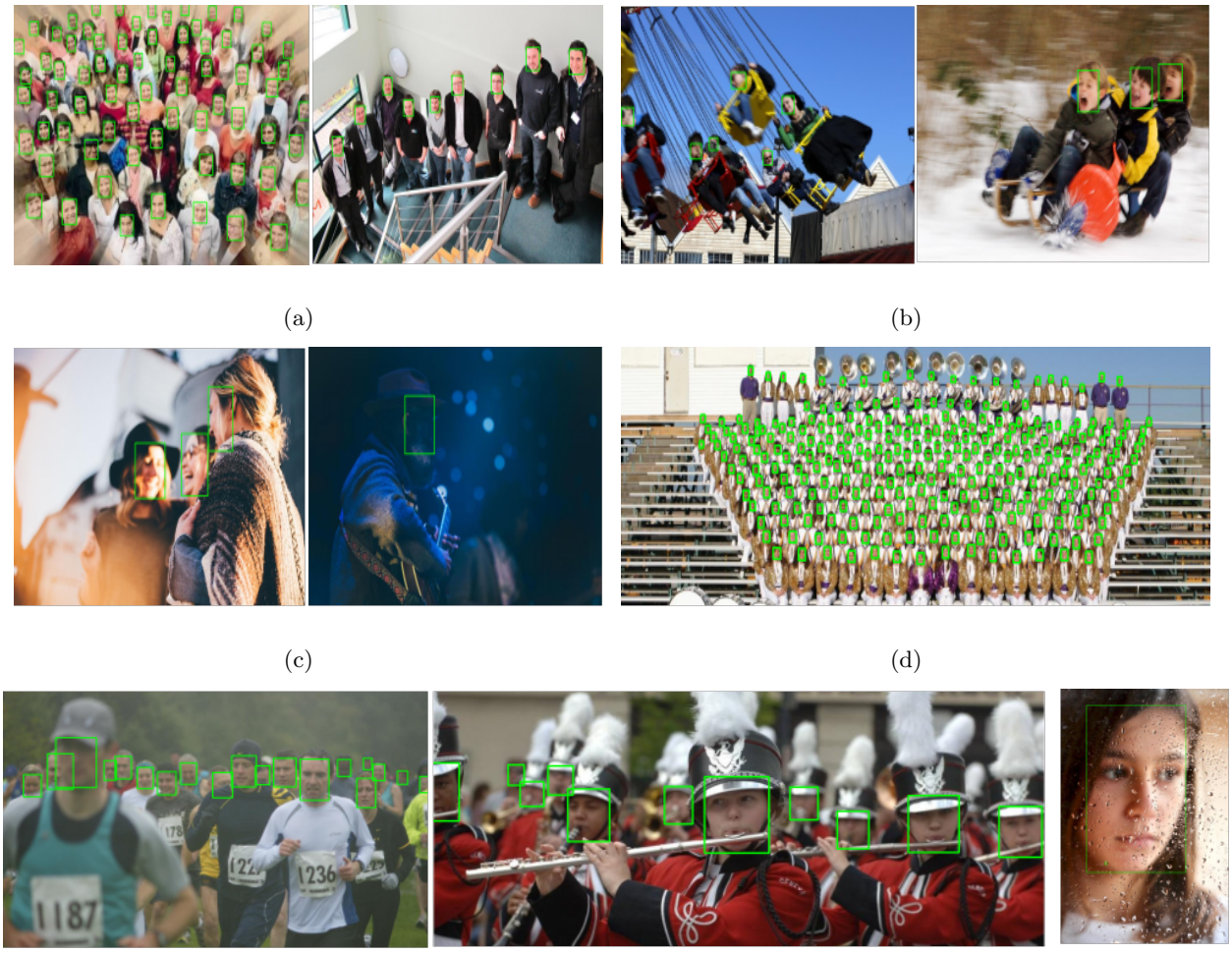
(c) (d)

Figure 3.23: Depiction of the feature maps. To investigate the influence of the gridding artifacts problem on the performance, feature maps with strides of 4 are extracted and visualized from different models. Each feature map is obtained by computing average values in the channel direction. The green boxes include the face regions. (a) Input image from the WIDER FACE dataset. (b) Depiction of the feature map generated by SANet. (c) Depiction of the feature map generated by SANet(Conv). (d) Depiction of the feature map generated by SANet(DConv). Due to loss of detailed facial information, facial features in the face regions In the analysis of the feature maps, it arises that the feature map depicted in (b) includes more discriminative and complete facial features than feature maps (c)(d) around the regions such as eyes, mouth, nose.

Table 3.16: mAP% results of the ablation study of the SANet on the UFDD dataset.

Methods	mAP%
S^3FD-M	69.1
S^3FD-M+SA	75.6
S^3FD-M+SA+SCEM	76.6
S^3FD-M+SA+SCEM+LB	77.0

Ablation Study: SANet To demonstrate the generalization ability of the AFFM and the SCEM on detecting faces with various scales in a complex environment, the ablation study of the SANet on the UFDD is conducted. Table 3.16 depicts the evaluation results of the AFFM and the SCEM on the UFDD datasets. Specifically, the AFFM with the spatial-attention module improves the performance from mAP of 69.1% to mAP of 75.6%, which demonstrates the effectiveness of the AFFM on detecting multi-scale faces in a complex environment. Furthermore, the SCEM improves the performance from a mAP of 75.6% to a mAP of 76.6%, which suggests the SCEM effectively enriches contextual information and improves the performance on top of the AFFM. Finally, SANet with a large batch size of 14 achieves the best performance with a mAP of 77.0. These observations demonstrated the generalization ability of the AFFM and the SCEM on finding various scales of faces in a complex environment and get the same conclusion as previous experiments that the AFFM and the SCEM improve the performance of detecting multi-scale faces consistently in a complex environment. Fig. 3.24 depicts selected qualitative results of SANet.



(a) (b) (c) (d) (e)

Figure 3.24: Qualitative results of the SANet under the different conditions. (a) full pose distribution; (b) motion blur; (c) illumination; (d) small faces (e) multi-scale faces in a complex environment.

4 Lessons Learned and Design Rules

Human Detection: In terms of human perception, occlusion could result in two problems. (1) The occlusion results in inadequate human information that is used to localize humans. (2) The large similarity between human appearance and background appearance distracts and confuses the network to distinguish the human and occluding objects. Therefore, if we want to improve performance, the above two problems should be mitigated by the designed network. Two-stage human detectors include RPN, R-CNN stages, and an RoI intermediate layer. The RPN stage performs coarse predictions, the R-CNN will refine the results of the RPN stage to get final predictions. Therefore, the improvements in the R-CNN stage could improve the final performance more than the improvements in the RPN stage. In general, the designed multi-branches or multi-task structure focuses on improvements on the R-CNN stage. Multi-task represents that the network is supervised by multiple supervision signals. Multi-branches represents that the network structure has multiple parallel network branches.

In the R-CNN stage, the input features are learned by RoI operation and based on the proposals produced by preset anchors. So, the features used for predictions in the R-CNN stage are region-wise features. Therefore, the contextual information of region-wise full-body features could be used as an effective way of localizing full-body positions. The contextual information represents the extra information around a bounding box and fused information. Therefore, fusing features learned by different supervision signals is an effective way of enriching contextual information, which has been demonstrated in our experiments. To increase the representation power of features learned by the network, the self-attention mechanism can be first considered for designing the structure of the neural network. Because the self-attention mechanism doesn't need extra supervision signals and extra annotations and improves performance. The essential of the attention mechanism is to learn different importance for reliable information and noise of a feature map. A designed module based on the self-attention mechanism often is used as an intermediate module within the network. Therefore, under the premise of improving performance, the number of convolutional layers and the

channels of convolutional layers should be set as less as possible. In terms of our experience, three convolutional layers with channels of 256, 128, and 1 are in line with our expectations of improved performance.

Learning discriminative features is difficult and always considered as an essential goal of improving performance. Detection has a serious contradiction that the goals of classification and regression tasks are different. The goal of the classification task is to learn translation-invariant features the goal regression task is to learn translation-variant features. However, the classification and regression tasks are performed in a unified network. Our experimental results suggest that the value of classification loss occupies most of the value of the multi-task loss. Therefore, improving the classification ability of the network could help the network to learn more discriminative features. To mitigate the influence of variant features for the classification task, the pixel-wise classification task was considered in our method and is an effective way of learning more discriminative features, which has been demonstrated in our experiments. Moreover, the designed module for pixel-wise classification tasks should have the same consideration as the designed module based on the attention mechanism. Finally, multi-task can provide more features of rich diversity for the network to perform a specific task. Moreover, multi-task often provides more supervision signals for the network to learn features, which means applying more constraints to the network. Therefore, multiple supervision signals could force the network to focus on reliable information and mitigate the problem of loss divergence. Therefore, with sufficient annotation resources, multiple supervision signals should be considered firstly because applying multiple supervision signals nearly doesn't increase the complexity of the model.

Face Detection: The common disadvantage of two-stage face detectors is constrained inference speed due to a more complex network structure than single-stage face detectors. Tiny faces with the variance of occlusion, pose, expression, and blur could lead to two problems in the neural network. (1) The loss of facial information results in inadequate facial information that is used to localize faces. (2) The deformed facial feature structure leads to the difficulty of learning discriminative facial features. The above problems should be considered when designing a network structure.

To provide adequate information on small faces for the network to learn features, feature maps at different levels are used to detect faces at various scales. Specifically, the high-resolution feature map is used to detect small faces. Because a high-resolution feature map maintains adequate information on small faces. The low-resolution feature map is used to detect medium and big faces. Because a low-resolution feature map includes enough facial information to detect faces. However, multi-scale feature maps increase the complexity of the network and lower inference speed. However, feature maps at different levels have different semantic levels. Although experimental results have indicated that multi-scale feature maps could improve performance, there is an unsolved problem in academia whether the network learns different features for faces at different scales. There is a hypothesis that facial features aren't changed with the variance of face scales in terms of human perception. Therefore, a single-scale feature map could be employed to learn multi-scale facial features and can maintain high inference efficiency. But that how to obtain the target single-scale feature map and how to choose the downsampling rate of the target single-scale feature map should be explored. In our experiments, the downsampling rates and combinations of feature maps at different levels have been considered as much as possible.

Due to the biometrics of face and hair, the contextual information of face could improve the performance of detecting faces in terms of human perception. But if the contextual information of the face is much larger than the face region, it could result in much noise which distracts the network to learn discriminative facial features because contextual information obtained by the network is closely related to receptive field size. Therefore, the designed feature fusion method should consider the receptive field size and scale of faces that are detected in a feature map with a specific size. Conservatively, the ratio of preset anchor scales and face scales that are needed to detect in a specific feature map ranges from 0.25 to 0.5, which avoids introducing much noise. When designing the context aggregation module, the kernel sizes should be considered. In general, the kernel sizes of different groups of convolutional layers range from 1×1, 3×3, or 5×5. If the kernel size of a convolutional layer is larger than 5×5, it will increase the complexity of the network. Therefore, if there are multiple groups of convolutional layers, convolutional layers could alternatively use a

kernel size of 1×1, 3×3, and 5×5 in one group. In terms of our experience, the kernel size of 1×1 and 3×3 are referred. Finally, if dilated convolutions are used to aggregate contextual information, the dilation rate could be considered instead of kernel size.

The unsolved problem is whether the network learns different features for faces at different scales in academia. To mitigate this problem, fusing features at different levels is a demonstrated effective way of improving performance. But the influence of fused features at different levels on the performance should be considered. After we observed that the values of feature maps at different levels, a module based on the self-attention mechanism was designed. Our experimental results have demonstrated that the designed module could reduce the noise in the fused features and improve performance. In terms of our experience, the values of feature maps and gradients can find problems and cues to solve the problems. Moreover, preset anchor scales should be carefully designed in terms of receptive field size and face scales. The design of preset anchor scales should consider the number of false positives. Because a large number of false positives could introduce the difficulty of training the network and loss divergence. In terms of our experience, when loss divergence is introduced or the predictions of the network deviate far from the ground truths, the proposals and feature maps learned by the network should be visualized to adjust whether the network predicts the bounding boxes in the correct direction. Moreover, the gradients and feature values could be outputted to check whether the variance of numerical values has abnormal or not. In our work, the face scales are set from 8×8 to 512×512 with two time steps. In the attention module, the last convolutional layer can't be fed into a relu activation layer. Because the output of a relu activation layer is equal or greater than 0. However, the output of a convolutional layer could be positive or negative. Therefore, the relu activation layer results in a reduced range of values of attention weights. In addition, the output of the designed module is not used for predictions directly. Because the output of the designed module has a different feature attribute from the features in the backbone network, which reflects in that the values of the output of the designed module have a large difference from the values of the convolutional layers that have the same depth as the output of the designed module. Therefore, a transformed layer that consists of one or two

convolutional layers is needed to maintain the consistency of the output of the designed module and the output of the backbone network in semantic levels.

The local connectivity of the convolutional layer allows the network to learn filters that maximally respond to a local region of the input, thus exploiting the spatial local correlation of the input. For an input image, a pixel is more correlated to the nearby pixels than to the distant pixels. A dilated convolution inserts zeros in the neighboring values of a convolutional kernel in terms of dilation rate. If the dilation rate is equal to 2, one 0 is inserted in the neighboring values of a convolutional kernel. If the dilation rate is equal to 3, two 0s are inserted in the neighboring values of a convolutional kernel. Therefore, when designing a module to solve the gridding artifacts problem, the complexity of the designed module should be considered. Because the complexity of the convolutional layer with a smaller kernel size is exponentially larger than the complexity of the convolutional layer with a larger kernel size. In our work, one dilated convolution layer is fed into one convolutional layer. Finally, if the number of groups of convolutional layers or dilated convolutions is more than 3, the loss value of the network decreases very slow. In our work, two groups of designed dilated modules are used to aggregate contextual information.

5 Future Work

Human Detection: To overcome the occlusion challenge, the following future work might be useful:

(i) *Relational Learning*: The heatmaps and qualitative results have suggested that the network tends to learn human features from a head supervision signal. Although our experimental results have demonstrated the head supervision signal is more powerful than visible-body and full-body supervision signals, it is still difficult for the network to detect pedestrians with significant occlusion. Therefore, understanding how to utilize the relationship among the head, visible-body, and full-body to infer the full-body positions will improve the performance of human detection.

(ii) *Explainable Features*: Two issues might result in constrained performance of human detection algorithms. First, the network couldn't employ sufficient visible-body information to localize full-body positions. Second, the variance of occluding object appearance distracts the network focusing on reliable information. Therefore, if the features learned by the network could be explainable, it might be easier to explain how the network addresses each issue.

(iii) *Blur in Near-infrared Images*: When blur occurs in the near-infrared images, our method tends to produce bounding boxes that have less overlap with the ground truth. Compared to near-infrared images with blur, the predictions of our method produce higher-quality bounding boxes on the near-infrared images without blur. Blur results in minor movements of pixels in the images. However, the goal of the classification task in the detection network is to learn translation-invariant features. Therefore, if this conflict could be resolved in future work, the performance will be improved. The near-infrared images are grayscale and thus have less color information than visible images. Moreover, capturing and annotating near-infrared images might be costlier than visible images. Therefore, if the visible images could be used to train the network and detect humans in the near-infrared images, it can contribute to speeding up

research on cross-domain detection problem.

Face Detection: To overcome the scale challenge, the following future work might be useful:

(i) *Receptive Field*: Magnifying the receptive field size is an effective way of enriching contextual information and improving performance on detecting multi-scale faces, which already has been demonstrated in many state-of-the-art face detectors. The receptive field consists of a theoretical receptive field and an effective receptive field. The theoretical receptive field is a theoretical region of the input image that is influenced by a pixel in the feature map. However, due to the characteristic of parameter updating of the network, the values in a region will not be updated equally. Therefore, the effective receptive field size is smaller than the theoretical receptive field size. Intuitively, if contextual information is too small around face anchor sizes, the network can't employ adequate contextual information to learn features. If contextual information is too large around face anchor sizes, it could provide much redundant information and much noise. Therefore, tiling more suitable face anchor sizes would improve performance. Because the face anchor sizes are set in terms of an effective receptive field, finding the ideal ratio of effective receptive field size and face anchor size will lead to the great progress of face detection.

(ii) *Feature Consistency*: With the decrease of face size, the facial appearance losses more information. Especially, when small faces have a variance of the pose, occlusion, expression, and blur, it is more difficult for the network to learn discriminative facial features. With the increase of the depth of the network, the downsampling will result in more loss of facial information. For single-stage face detectors, to provide adequate facial information for the network to learn features, single-stage face detectors employ feature maps at different levels to detect faces at various scales. However, the features maps at different levels produce semantic information at different levels. Therefore, how the network treats the face at various scales should be explored in future work.

(iii) *Working Mechanism*: In general, two-stage object detectors perform better than single-stage

object detectors due to the refinement of the second stage. However, the qualitative results of the WIDER FACE dataset indicate that single-stage face detectors are dominant in the face detection benchmark. Therefore, understanding how the second stage works in the two-stage face detector should be explored. Moreover, a comparative study should be conducted to validate whether the RoI layer has a large influence on performance or not.

6 Conclusions

This dissertation focused on designing, developing, and evaluating human detection algorithms in the wild. The primary contributions were achieved by developing human detector and face detectors to overcome occlusion and scale challenges, respectively.

A human detector was developed that was robust to occlusion variations. Detailed experiments were conducted on the CityPersons dataset to demonstrate that the proposed human detector was robust to detect humans at different occlusion levels. The proposed human detector achieved the same results as the best performance of compared methods and requires fewer parameters. Specifically, our designed AFIM is the first to use feature interaction between multi-branches network for detecting occluded humans. The specialty of our designed module is that feature interaction was applied to a two-branches network that is supervised by different supervision signals. So our designed module provided an insight into enriching contextual information of features learned by different supervision signals. Our qualitative and quantitative results demonstrated that the designed feature interaction module helps the network to learn more discriminative human features by increasing the contrast of values of human features and background features. Moreover, detailed experiments were conducted to demonstrate the effectiveness of the pixel-wise classification task on improving performance. Through visualization of heatmaps of features learned by pixel-wise classification module, we observed that the designed pixel-wise classification module increased the response of human features, but also increase the response of background features. However, the difference of values of human features and background features learned by the designed module is still larger than that of human and background features learned by the network without the designed module. The qualitative and quantitative results demonstrated that the designed pixel-wise classification module enhances the sensitivity of the network to human features at the pixel level. Therefore, the pixel-wise classification task could be used as an implement of a region-wise classification task to enhance the sensitivity of the network to object position at the pixel level, which improves the performance of the regression task. Finally, our experimental results indicated

a surprising observation that a head supervision signal is more powerful than visible-body and full-body supervision signals. It leads to that the features learned by the head, visible-body, full-body supervision signals are only around the head region. The observation suggests that the network tends to distinguish humans and backgrounds through learning features from the head region and discriminative information of humans. This is conducive to the effective use of local features on inferring global structure in subsequent research.

To overcome the scale challenge, multi-scale feature maps are used as an effective way of learning discriminative facial features for improving performance. However, multi-scales feature maps result in high complexity of the network and slow inference speed compared to a single-scale feature map. Two-stage face detectors also lead to this issue compared to single-stage face detectors. Therefore, a comparative study was conducted to investigate which feature map at a specific level or combination of feature maps at different levels contributes more to improving the performance of detecting multi-scale features. Our experimental results demonstrated the feature map obtained by a deep convolutional layer contributes more than the feature map obtained by a shallow convolutional layer and their combinations. To maintain adequate information of small faces, the transposed convolutional layer was used to magnify the feature map obtained by the last convolutional layer of the backbone network. Our qualitative and quantitative results demonstrated that a single-scale feature map is an effective way of learning multi-scale features. To address the problem that small faces result in inadequate information, obtaining contextual information around the face region by magnifying receptive field size is needed, which is in line with human perception. Adding convolutional layers or combining features from different receptive field sizes could magnify receptive field size. However, adding convolutional layers could increase the depth of the network and increase the difficulty of training the network. Therefore, we designed a context aggregation module that combines features obtained by two groups of dilated convolutional layers. Our experimental results have demonstrated contextual information can improve performance.

Due to the high degree of extensibility and inference performance of single-stage face detectors, multi-scale feature maps are always applied to single-stage face detectors. In general, due

to the variance of semantic levels of feature maps at different levels, the combination of feature maps at different levels is considered by all state-of-the-art single-stage face detectors. Through our experiments, we observed that the numerical values of feature maps at different levels have a large difference. Therefore, noise in different feature maps will occupy different proportions of noise in the fused feature map. Our experimental results indicated that noise in the feature map obtained by a deep layer occupies most of the noise in the fused feature map. Therefore, to reduce noise that could distract the network to focus on reliable facial information, we designed a module to reduce noise in the fused feature map by applying the self-attention mechanism to high-level semantic feature map. Because the attention mechanism learns different importance values for face region and background, we hypothesized the attention mechanism can be used for reducing noise and improving performance. Our experimental results indicated that the self-attention mechanism applied to high-level feature maps and fusing features at different levels could reduce noise and improving performance. Moreover, a comparative study was conducted to investigate the influence of channel-wise and spatial-wise attention mechanisms on performance. The experimental results demonstrated that spatial-wise information works better than channel-wise information for detecting multi-scale faces while requiring fewer parameters. Although dilated convolutions could magnify receptive field size, dilated convolutions could result in the gridding artifacts problem. Our designed module is the first to solve the gridding artifacts problem for face detection. Our designed module altered dilated convolutions and convolutional layers and combined features from different receptive field sizes. The convolutional layer was used to re-learn the relationship between separable units obtained by dilated convolution to solve the gridding artifacts problem. Our experimental results demonstrated our method is easy and effective in improving performance. Also, we observed that although dilated convolutions can lead to the gridding artifacts problem, only using dilated convolutions performs better than only using convolutional layers. It suggests that contextual information can mitigate the gridding artifacts problem. This is conducive to investigating how the network treats the noise and reliable information in the subsequent research.

Bibliography

[1] Abdelgader, A., and Viriri, S. A survey on soft biometrics for human identification. *Machine Learning and Biometrics 1* (2018), 37.

[2] Appel, R., Fuchs, T., Dollár, P., and Perona, P. Quickly boosting detection trees-pruning underarchieving features early. In *Proceedings of International Conference on Machine Learning* (Atlanta, GA, 2013), pp. 594–602.

[3] Bottou, L. Stochastic gradient descent tricks. *Neural Networks: Tricks of the Trade 7700* (2012), 421–436.

[4] Cai, Z., and Vasconcelos, N. Cascade R-CNN: delving into high quality object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT, 2018), pp. 6154–6162.

[5] Chen, L., George, P., Iasonas, K., Kevin, M., and Alan, L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *Proceedings of IEEE Conference on Learning Representations* (San Diego, CA, 2015), pp. 1–10.

[6] Chen, Y., Han, C., Li, Y., Huang, Z., Jiang, Y., Wang, N., and Zhang, Z. SimpleDet: A simple and versatile distributed framework for object detection and instance recognition. *Journal of Machine Learning Research 20* (2019), 1–8.

[7] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscaps dataset for semantic urban scene understanding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV, 2016), pp. 3213–3223.

[8] Dollár, P., Wojek, C., Schiele, B., and Perona, P. Pedestrian detection: A benchmark. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Miami Beach, FL, 2009), pp. 304–311.

[9] Dou, P., Shah, S. K., and Kakadiaris, I. A. End-to-end 3D face reconstruction with deep neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI, 2017), pp. 5908–5917.

[10] Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Rhode, Island, 2012), pp. 3354–3361.

[11] Girshick, R. Fast R-CNN. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA, 2015), pp. 1440–1448.

[12] Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH, 2014), pp. 580–587.

[13] Glorot, X., and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of International Conference on Artificial Intelligence and Statistics* (Sardinia, Italy, 2010), pp. 249–256.

[14] He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask R-CNN. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI, 2017), pp. 2961–2969.

[15] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV, 2016), pp. 770–778.

[16] Hu, P., and Ramanan, D. Finding tiny faces. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI, 2017), pp. 951–959.

[17] Jain, V., and Learned-Miller, E. FDDB: A benchmark for face detection in unconstrained settings. *UMass Amherst Technical Report 2* (2010), 5.

[18] Le, H., Smailis, C., Shi, L., and Kakadiaris, I. A. EDGE20: A cross spectral evaluation dataset for multiple surveillance problems. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision* (Snowmass village, CL, 2020), pp. 2685–2694.

[19] Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., and Huang., F. DSFD: dual shot face detector. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA, 2019), pp. 5060–5069.

[20] Li, Q., Jin, S., and Yan, J. Mimicking very efficient network for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI, 2017), pp. 6356–6364.

[21] Li, Y., Sun, B., Wu, T., and Wang, Y. Face detection with end-to-end integration of a ConvNet and a 3D model. In *Proceedings of European Conference on Computer Vision* (Amsterdam, Netherlands, 2016), pp. 420–436.

[22] Li, Z., Tang, X., Han, J., Liu, J., and He, R. PyramidBox++: High performance detector for finding tiny face. *CoRR 1* (2019), 1–10.

[23] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie., S. Feature pyramid networks for object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI, 2017), pp. 2117–2125.

[24] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollar, P. Focal loss for dense object detection. In *Proceedings of IEEE International Conference on Computer Vision* (Venice, Italy, 2017), pp. 2980–2988.

[25] Liu, L., Li, G., Xie, Y., Yu, Y., Wang, Q., and Lin., L. Facial landmark machines: A backbone-branches architecture with progressive representation learning. *IEEE Transactions on Multimedia 21* (2019), 790–799.

[26] Liu, S., Di, H., and Wang., Y. Adaptive NMS: Refining pedestrian detection in a crowd. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA, 2019), pp. 6459–6468.

[27] Liu, S., Huang, D., and Wang, Y. Receptive field block net for accurate and fast object detection. In *Proceedings of European Conference on Computer Vision* (Munich, Germany, 2018), pp. 385–400.

[28] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., and Berg, A. C. SSD: single shot multibox detector. In *Proceedings of European Conference on Computer Vision* (Amsterdam, Netherlands, 2016), pp. 21–37.

[29] Liu, W., Liao, S., Hu, W., Liang, X., and Chen, X. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of European Conference on Computer Vision* (Munich, Germany, 2018), pp. 618–634.

[30] Liu, W., Liao, S., Ren, W., Hu, W., and Yu, Y. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA, 2019), pp. 5187–5196.

[31] Luo, W., Li, Y., Urtasun, R., and Zemel, R. Understanding the effective receptive field in deep convolutional neural networks. In *Proceedings of International Conference on Neural Information Processing Systems* (Barcelona, Spain, 2016), pp. 4905–4913.

[32] Mathias, M., Benenson, R., Pedersoli, M., and Gool, L. V. Face detection without bells and whistles. In *Proceedings of European Conference on Computer Vision* (Zurich, Switzerland, 2014), pp. 720–735.

[33] Najibi, M., Samangouei, P., Chellappa, R., and Davis, L. S. SSH: single stage headless face detector. In *Proceedings of IEEE International Conference on Computer Vision* (Venice, Italy, 2017), pp. 4875–4884.

[34] Noh, J., Lee, S., Kim, B., and Kim, G. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Lake City, UT, 2018), pp. 966–974.

[35] Pang, Y., Xie, J., Khan, M. H., Anwer, R. M., Khan, F. S., and Shao, L. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of IEEE International Conference on Computer Vision* (Seoul, Korea, 2019), pp. 4967–4975.

[36] Pedagadi, S., Orwell, J., Velastin, S., and Boghossian, B. Local fisher discriminant analysis for pedestrian re-identification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Portland, OR, 2013), pp. 3318–3325.

[37] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV, 2016), pp. 779–788.

[38] Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of International Conference on Neural Information Processing Systems* (Stockholm, Sweeden, 2015), pp. 91–99.

[39] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision 115* (2015), 211–252.

[40] Samangouei, P., Najibi, M., Davis, L., and Chellappa, R. Face-Magnet: magnifying feature maps to detect small faces. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision* (Lake Tahoe, NV, 2018), pp. 122–130.

[41] Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., and Sun, J. CrowdHuman: a benchmark for detecting human in a crowd. *CoRR 1* (2018), 1–10.

[42] Shrivastava, A., Gupta, A., and Girshick, R. Training region-based object detectors with online hard example mining. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV, 2016), pp. 761–769.

[43] Song, T., Sun, L., Xie, D., Sun, H., and Pu, S. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *Proceedings of European Conference on Computer Vision* (2018), pp. 536–551.

[44] Tang, X., Du, D. K., He, Z., and Liu, J. PyramidBox: A context-assisted single shot face detector. In *Proceedings of European Conference on Computer Vision* (Munich, Germany, 2018), pp. 797–813.

[45] Tian, W., Wang, Z., Shen, H., amd Binghui Chen, W. D., and Zhang, X. Learning better features for face detection with feature fusion and segmentation supervision. *CoRR 1* (2018), 1–10.

[46] Triantafyllidou, D., Nousi, P., and Tefas., A. Fast deep convolutional face detection in the wild exploiting hard sample mining. *Big data research*, 11 (2018), 65–76.

[47] Wang, H., Li, Z., Ji, X., and Wang, Y. Face R-CNN. *CoRR 1* (2017), 1–10.

[48] Wang, J., Yuan, Y., and Yu, G. Face attention network: An effective face detector for the occluded faces. *CoRR 2* (2017), 1–10.

[49] Wang, X., Xiao, T., Jiang, Y., Shao, S., Sun, J., and Shen, C. Repulsion loss: detecting pedestrians in a crowd. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Lake City, UT, 2018), pp. 7774–7783.

[50] Wang, Y., Ji, X., Zhou, Z., Wang, H., and Li, Z. Detecting faces using region-based fully convolutional networks. *CoRR 3* (2017), 1–10.

[51] Wang, Z., and Ji, S. Smoothed dilated convolutions for improved dense prediction. In *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining* (London, United Kingdom, 2018), pp. 2486–2495.

[52] Xu, X., and Kakadiaris, I. A. Joint head pose estimation and face alignment framework using global and local CNN features. In *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition* (Washington, DC, 2017), pp. 642–649.

[53] Xu, X., Le, H., Dou, P., Wu, Y., and Kakadiaris, I. A. Evaluation of 3D-aided pose invariant 2D face recognition system. In *Proceedings of International Joint Conference on Biometrics* (Denver, CO, 2017), pp. 446–455.

[54] Yan, J., Zhang, X., Lei, Z., and Li., S. Z. Face detection by structural models. *Image and Vision Computing 32*, 10 (2017), 790–799.

[55] Yang, S., Luo, P., Loy, C. C., and Tang, X. WIDER FACE: A face detection benchmark. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV, 2016).

[56] Yang, S., Luo, P., Loy, C. C., and Tang, X. Faceness-net: Face detection through deep facial part responses. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40* (2017), 1845–1859.

[57] Yang, S., Xiong, Y., Loy, C. C., and Tang, X. Face detection through scale-friendly deep convolutional networks. *CoRR 5* (2017), 1–10.

[58] Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. Understanding neural networks through deep visualization. *CoRR 1* (2015), 1–10.

[59] Yu, F., and Koltun, V. Multi-scale context aggregation by dilated convolutions. *CoRR 2* (2015), 1–10.

[60] Yu, W., Yang, K., Bai, Y., Xiao, T., Yao, H., and Rui, Y. Visualizing and comparing Alexnet and VGG using deconvolutional layers. In *Proceedings of International Conference on Machine Learning* (Colorado Springs, CO, 2016).

[61] Z, S., and Li. Markov random field models in computer vision. In *Proceedings of European Conference on Computer Vision* (Stockholm, Sweden, 1994), pp. 361–370.

[62] Zhang, J., Wu, X., Zhu, J., and Hoi, S. C. Feature agglomeration networks for single stage face detection. *Neurocomputing 380* (2020), 180–189.

[63] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters 23*, 10 (2016), 1499–1503.

[64] Zhang, K., Zhang, Z., Wang, H., Li, Z., Qiao, Y., and Liu., W. Detecting faces using inside cascaded contextual cnn. In *Proceedings of IEEE International Conference on Computer Vision* (Venice, Italy, 2017), pp. 3171–3179.

[65] Zhang, L., Lin, L., Liang, X., and He, K. Is Faster R-CNN doing well for pedestrian detection? In *Proceedings of European Conference on Computer Vision* (Amsterdam, Netherlands, 2016), pp. 443–457.

[66] Zhang, S., Benenson, R., and Schiele, B. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI, 2017), pp. 3213–3221.

[67] Zhang, S., Wen, L., Lei, Z., and Li, S. Z. Occlusion-aware R-CNN: detecting pedestrians in a crowd. In *Proceedings of European Conference on Computer Vision* (Munich, Germany, 2018), pp. 637–653.

[68] Zhang, S., Wen, L., Shi, H., Lei, Z., Lyu, S., and Li., S. Z. Single-shot scale-aware network for real-time face detection. *International Journal of Computer Vision 127* (2018), 537–559.

[69] Zhang, S., Yang, J., and Schiele, B. Occluded pedestrian detection through guided attention in CNNs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Lake City, UT, 2018), pp. 6995–7003.

[70] Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., and Li, S. Z. S^3FD: single shot scale-invariant face detector. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI, 2017).

[71] Zhou, C., and Yuan, J. Bi-box regression for pedestrian detection and occlusion estimation. In *Proceedings of European Conference on Computer Vision* (Munich, Germany, 2018), pp. 135–151.

[72] Zhu, C., Tao, R., Luu, K., and Savvides, M. Seeing small faces from robust anchor's perspective. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT, 2018), pp. 5127–5136.

[73] Zhu, X., and Ramanan, D. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (RI, 2012), pp. 2879–2886.

[74] Zitnick, C. L., and Dollar, P. Edge boxes: Locating object proposals from edges. In *Proceedings of European Conference on Computer Vision* (Zurich, Switzerland, 2014), pp. 1–10.