Summer 2022

Image Restoration Under Adverse Illumination for Various Applications

Lan Fu

Image restoration under adverse illumination for various applications

By

Lan Fu

Bachelor of Engineering
Yanshan University, 2011

Master of Engineering
Tianjin University, 2014

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Computer Science and Engineering

College of Engineering and Computing

University of South Carolina

2022

Accepted by:

Song Wang, Major Professor

Michael N. Huhns, Committee Member

Yan Tong, Committee Member

Lannan Luo, Committee Member

Hongkai Yu, Committee Member

Tracey L. Weldon, Vice Provost and Dean of Graduate Studies

Acknowledgments

At first, I would like to thank my advisor Prof. Song Wang. Dr. Wang supervised me under this doctoral thesis by providing significant help. His profound knowledge and rigorous scholarship drive me to keep diving into computer vision and overcome the difficulties in research. His tremendous help and support guide me through the difficulties in my research. His high academic standard also encourages me to be a better researcher.

I am truly grateful to my colleagues Dr. Hongkai Yu and Dr. Qing Guo. They have insightful and sharp sense of problems, which keeps me staying in the right direction in my research. Their self-motivated personality inspires me a lot, especially when I struggle with difficulties in my research. We cooperate with each other and discuss a lot about the research projects as a team, and this experience teaches me how to do a solid research work, how to formulate a problem systematically, and how to broaden my minds.

I would also like to thank my dissertation committee members, Prof. Michael Huhns, Prof. Yan Tong, and Prof. Lannan Luo, for their help and valuable suggestions on my work. It is my great honor to have them as my committee members.

I am truly grateful to my colleagues and friends Dr. Jeff Simmons, Dr. Felix-Juefei Xu, Changqing Zhou, Xinyi Wu, Zhenyao Wu and other fellow labmates. Open-minded discussion with them always inspire me a lot in my research, and the inspiring and friendly environment makes me very impressive and grateful.

Finally, I want to express my deepest appreciation to my family. Their endless love is the best comfort to support me in the journey of pursuing my Ph.D. degree.

Abstract

Many images are captured in sub-optimal environment, resulting in various kinds of degradations, such as noise, blur, and shadow. Adverse illumination is one of the most important factors resulting in image degradation with color and illumination distortion or even unidentified image content. Degradation caused by the adverse illumination makes the images suffer from worse visual quality, which might also lead to negative effects on high-level perception tasks, *e.g.*, object detection.

Image restoration under adverse illumination is an effective way to remove such kind of degradations to obtain visual pleasing images. Existing state-of-the-art deep neural networks (DNNs) based image restoration methods have achieved impressive performance for image visual quality improvement. However, different real-world applications require the image restoration under adverse illumination to achieve different goals. For example, in the computational photography field, visually pleasing image is desired in the smartphone photography. Nevertheless, for traffic surveillance and autonomous driving in the low light or nighttime scenario, high-level perception tasks, *e.g.*, object detection, become more important to ensure safe and robust driving performance. Therefore, in this dissertation, we try to explore DNN-based image restoration solutions for images captured under adverse illumination in three important applications: 1) image visual quality enhancement, 2) object detection improvement, and 3) enhanced image visual quality and better detection performance simultaneously.

First, in the computational photography field, visually pleasing images are desired. We take shadow removal task as an example to fully explore image visual

quality enhancement. Shadow removal is still a challenging task due to its inherent background-dependent and spatial-variant properties, leading to unknown and diverse shadow patterns. We propose a novel solution by formulating this task as an exposure fusion problem to address the challenges. We propose shadow-aware FusionNet to 'smartly' fuse multiple over-exposure images with pixel-wise fusion weight maps, and boundary-aware RefineNet to eliminate the remaining shadow trace further. Experiment results show that our method outperforms other CNN-based methods in three datasets.

Second, we explore the application of CNN-based night-to-day image translation for improving vehicle detection in traffic surveillance that is important for safe and robust driving. We propose a detail-preserving method to implement the nighttime to daytime image translation and thus adapt daytime trained detection model to nighttime vehicle detection. We utilize StyleMix method to acquire paired images of daytime and nighttime for the nighttime to daytime image translation training. The translation is implemented based on kernel prediction network to avoid texture corruption. Experimental results showed that the proposed method can better address the nighttime vehicle detection task by reusing the daytime domain knowledge.

Third, we explore the image visual quality and facial landmark detection improvement simultaneously. For the portrait images captured in the wild, the facial landmark detection can be affected by the cast shadow. We construct a novel benchmark SHAREL covering diverse face shadow patterns with different intensities, sizes, shapes, and locations to study the effects of shadow removal on facial landmark detection. Moreover, we propose a novel adversarial shadow attack to mine hard shadow patterns. We conduct extensive analysis on three shadow removal methods and three landmark detectors. Then, we design a novel landmark detection-aware shadow removal framework, which empowers shadow removal to achieve higher restoration quality and enhances the shadow robustness of deployed facial landmark detectors.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

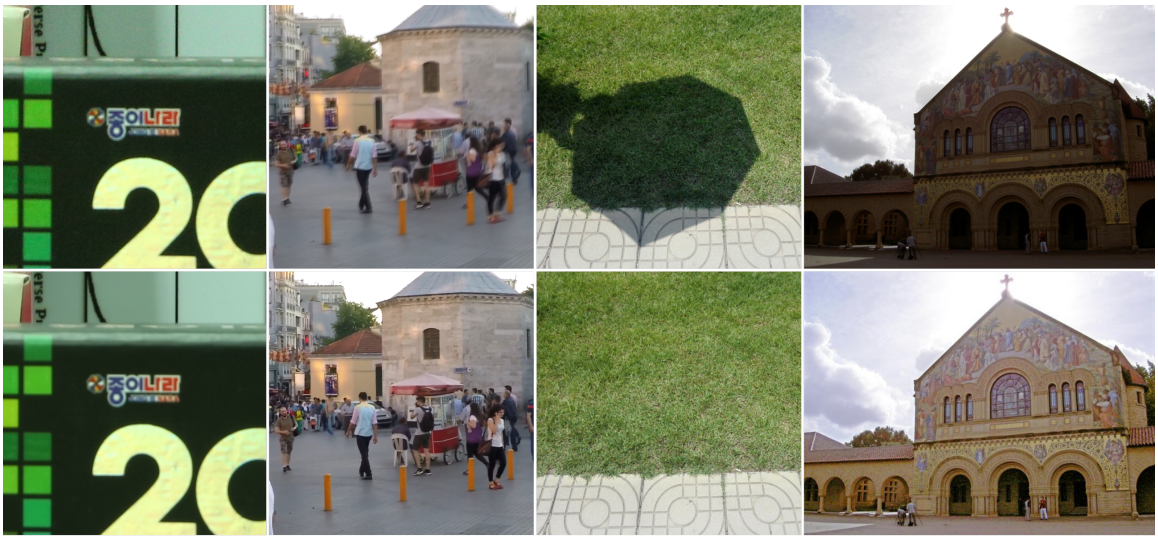

Figure 1.1: Illustration of image restoration tasks: denoising, deblurring, shadow removal and low-light image enhancement. The first row indicates the noisy, blur, shadow and low-light images, while the second row represents the corresponding predicted degradation-free images after image restoration.

Images are often captured under sub-optimal environments, *e.g.*, adverse illumination, or inappropriate camera settings, bringing different kinds of degradations to images, such as noise, blur, shadow, and low-light degradation, as shown in the first row of Fig. 1.1. Such kinds of degradations hurt the image aesthetic quality significantly. As a result, the appearance of degraded images lacks fine details, such as reduced contrast and inaccurate color, even worse, the content of the degraded images are not recognizable. For example, the shape and edge information no longer remains salient in the blur image shown in Fig. 1.1. Thus, low-quality image does not achieve satisfactory transmission of information, suffering from unpleasing visual experience.

Due to complex environment, insufficient illumination is one of the most common factor causing image degradation, receiving substantial attention in computer vision community [53, 69, 26, 36, 59]. Images taken under sub-optimal illumination, such as

backlit, uneven light, dim light, or where light source is blocked by an external object, suffer from severe visual quality degradation. A shadow image is generated where light source is blocked. A low-light image or a nighttime image is collected where images are captured under the dim light or extremely dark environment. Adverse illumination introduces color and illumination distortion into images, as shown by the shadow image in Fig. 1.1. Even worse, the scene content disappears in the degraded image, making the objects inside the image hard to recognize for human eyes, as shown in the low-light image in Fig. 1.1.

Image restoration is an effective way to restore a high-quality image by modeling the degradation pattern to achieve visually pleasing effect. Given the four types of degradations mentioned above, the corresponding image restoration tasks are denoising, deblurring, shadow removal, and low-light image enhancement. In this dissertation, we mainly focus on such image restorations under adverse illumination because it is prevalent in real-world applications and brings severe degradation to images. Image restoration under adverse illumination aims to recover a visually pleasing image without noise and with accurate color and illumination. Taking the shadow removal task for example, given a shadow image as input, the shadow removal algorithm predicts a shadow-free image, as shown in Fig. 1.1.

Early image restoration methods [123, 27, 43, 143, 106, 129] under adverse illumination usually model the illumination prior to achieve high-quality images with clear details. However, these methods have limited feature representation ability due to the use of handcrafted features. Recently, due to the rich feature representation learning ability, deep learning based image restoration models targeting at adverse illumination have achieved state-of-the-art (SOTA) performance to obtain high-quality images with pleasing visual experience without obvious noise and color distortion. Specifically, deep learning based shadow removal methods [53, 69] perform shadow removal from the perspective of image-to-image translation and physical shadow model with

unpaired and paired shadow and shadow-free images, respectively. Low-light image enhancement [36] or night-to-day image translation can perform image restoration to get a high-quality image with more and clearer details.

1.2 Motivation

Image restoration under adverse illumination enjoys a wide range of applications in the real world, including visual surveillance, autonomous driving, and computational photography. Current image restoration algorithms usually evaluate their performance based on the perceptual visual quality in the form of objective evaluation metrics, such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). However, diverse real-world applications demand different goals for the image restoration task under adverse illumination.

First, computational photography has become ubiquitous and prominent. Taking photographs in the wild is especially challenging due to the complicated environment. Shadow or low-light images are captured when the illumination is adverse. Different from low-light image enhancement, the shadow removal task is to recover the color and illumination in the shadow region and preserve the details in the non-shadow region as well to acquire a visual-pleasing image. Note that the shadow degradation presents background-dependent and spatial-variant properties, as shown in Fig. 1.2(a), making image restoration algorithms hard to capture its pattern. Specifically, the contiguous shadow cast on the background image may cause the shadow region to appear differently based on how the original shadow-free background region looks like, as well as where the shadow is cast on the background image in a pixel-wise manner. Moreover, due to the partially shadowed region along the shadow boundary, it is still challenging for the current SOTA shadow removal algorithms to obtain traceless background. All of these mentioned above motivate us to design an opti-

Figure 1.2: Illustration of (a) a paired shadow and shadow-free images, (b) an unpaired nighttime and daytime images, where red bounding boxes indicates the vehicle detection results by the pretrained daytime vehicle detector and detection performance on the nighttime image decreases, and (c) a paired shadowed face and clean face images, where green dots indicate the ground-truth facial landmarks while the red dots are the predicted facial landmarks by the pretrained facial landmark detector on clean images. The landmark detection performance drops on the shadowed face image.

mal shadow removal model to better capture the shadow pattern and obtain visually pleasing image with higher image quality.

Second, for traffic surveillance and autonomous driving in the low light or nighttime scenario, high-level perception tasks, *e.g.*, object detection and semantic segmentation, become more important to ensure safe and robust driving experience. Specifically, deep learning based object detectors play a vital role in identifying and localizing objects, such as pedestrians, vehicles, traffic signs, and barriers in the vehicle's vicinity. However, image captured under adverse illumination degrades the image quality, and generally the extracted features from the degraded image become corrupted as well, *i.e.*, image degradation leads to a domain distribution shift between clean and degraded images. As a result, domain gap will lead to a performance drop when a deep model pretrained on clean images is evaluated on degraded images.

For example, the cars' appearances, *e.g.*, edge and shape, in the nighttime image in Fig. 1.2(b), become diffused due to the light reflection and light scattering under the poor illumination. It might be hard for vehicle detector trained on daytime images to identify the degraded vehicles. Although image restoration for adverse illumination plays an important role to obtain a high-quality image, its application on high-level object detection task is not well studied. An intuitive way to alleviate the detection performance drop caused by domain shift is that we can transfer the nighttime image to its daytime version utilizing the image restoration method, then the detection performance may be improved.

Third, in some real-world applications, high-quality visual experience and accurate object detection are equally important goals to achieve for image restoration under adverse illumination. For example, portrait facial images captured in the wild are easily covered by shadow because of the complicated environment where images are collected, which hurts the image visual quality and the facial landmark detection performance as well, as shown in Fig. 1.2(c). In such scenario, pleasing visual experience as well as the accurate landmark detection are desired at the same time, because the former conveys the main message about the subject in the image and the latter is a fundamental step for facial related tasks, such as face recognition and facial expression recognition. Thus, linking the image restoration and object detection tasks are meaningful to explore. An intuitive way to alleviate the performance loss caused by shadow is to restore the underlying shadow-free image utilizing current SOTA shadow removal algorithms. Nevertheless, there are two challenges posing to such a solution: ① The interplay between light, occluder, and the subject directly affects the shadow appearance. As a result, in the real world, shadow patterns are significantly diverse, which increases the difficulty of shadow removal algorithms. ② Even though shadow removal methods could obtain high visual-quality images with lower RMSE, the landmark detection performance may even get worse compared to

that of shadow images due to the potential domain shift between landmark detection and image quality enhancement. Existing works (haze [94] and rain [50] removal) demonstrate that visual quality improvement benefits little or even hurts the high-level perception task performance. All above facts motivate us to answer two basic questions: how shadow affects the landmark detection, and whether shadow removal can benefit the robustness of landmark detectors or not. To this end, we propose to link the two seemingly independent but intrinsically related tasks, *i.e.*, shadow removal and facial landmark detection, by constructing a totally novel dataset and benchmark.

To summarize, in this dissertation, we focus on proposing CNN-based image restoration solutions under adverse illumination from the requirements of different real-world applications to achieve ① image visual quality enhancement, ② object detection improvement, and ③ enhanced image visual quality and better detection performance simultaneously.

1.3 Scope of the Proposed Research

This dissertation aims to explore the image restoration under adverse illumination in three applications, we summarize them as:

- Image visual quality enhancement. We take shadow removal task as an example to fully explore the image quality enhancement gain. Shadow casts where the light source is blocked. Shadow regions are usually darker due to insufficient illumination, presenting background-dependent and spatial-variant properties. Shadow removal task aims to recover the shadow-free image given the shadow image as input.
- Object detection performance improvement. We take nighttime vehicle detection as an example to reduce the performance lost caused by the undesired illumination. A pre-trained vehicle detector on daytime images suffers from performance

decrease when testing on nighttime images. We translate the nighttime image to its daytime version, and reuse the daytime vehicle detection model on the translated images to verify whether the performance will be improved.

- Both image visual quality and object detection improvement. We take shadow removal and facial landmark detection tasks as example to explore whether enhanced image quality obtained by shadow removal task can also achieve better facial landmark detection performance simultaneously.

For the image visual quality improvement, we propose a novel method, named auto-exposure fusion network, for single image shadow removal. We first utilize exposure estimation to learn multiple over-exposure images by compensating the shadow region with different exposure levels. Then we propose the *shadow-aware FusionNet* to produce fusion weight maps across all the over-exposure images. It can 'smartly' select which over-exposed pixel is the best one to recover the position-specific background. The proposed method fuses the input image and its over-exposure versions in a pixel-wise way. Further, we propose a *boundary-aware RefineNet* to remove the remaining shadow trace for refining the removal result obtained in the previous step. The proposed methods can obtain traceless background image than the state-of-the-art methods SP+M-Net [69] and DSC [52].

For the object detection improvement, we choose the vehicle detection problem in the traffic surveillance video as a case study and would like to reuse the daytime perception model to nighttime scenarios. Our basic idea is to maximally use the pre-trained daytime perception model, which could be easily extended to the nighttime tasks. Reversely to the traditional methods, we transfer the nighttime images back to the daytime style with the detail-preserving to reuse the trained daytime perception model. The strengths of this reverse way are obvious and promising: 1) there are no extra training efforts for the already trained daytime perception model and no needs to manually label the nighttime data; 2) image transfer could reduce the domain

distribution discrepancy between daytime and nighttime data; 3) detail-preserving image transfer could better maintain the structure details than the generative adversarial network (GAN) based image transfer. GAN is a general way to perform image translation. However, because there are downsampling and upsampling operations in a deep generator, which make GAN suffers from possible losing some structure details. Specifically, we propose a detail-preserving unpaired domain transfer method for this task, which mainly contains two components: 1) Style-transfer based StyleMix, 2) Kernel Prediction Network (KPN) based nighttime to daytime image transfer. Without paired daytime-nighttime image pairs, we propose to utilize the style translation based StyleMix method, inspired by AugMix [49], to acquire pairs of daytime and nighttime images as training data for the following nighttime to daytime image transfer. We can effectively alleviate the detail corruption caused by GAN: ① The synthetic nighttime image and corresponding daytime image translation can provide pixel-wise correspondence for night-to-day translation. ② Kernel prediction network based method can refine the nighttime to daytime image translation because the per-pixel kernel fusion can effectively utilize the neighboring region for each pixel and could learn more spatial context representing structure information. The proposed method can conduct daytime and nighttime vehicle detection with just one daytime model, which is more convenient in real-world applications.

For the image visual quality and object detection improvement simultaneously, we study whether shadow removal can improve the robustness of facial landmark detectors by constructing a totally novel dataset and benchmark. The constructed benchmark SHAREL covers synthetic data, adversarial data, and real data to quantitatively and systematically study the effects of shadow and shadow removal on the facial landmark detection. The synthetic subset mainly constructs shadow data based on the physical shadow model and covers diverse shadow patterns by modeling the physical shadow parameters. The shadow patterns consider four factors (*i.e.*, inten-

sity, size, shape, and location) with three severities. The adversarial data is a novel adversarial shadow attack method which poses great challenges to current landmark detectors. We also verify shadow removal models on a real-world shadow face dataset for better generalization ability. We cover three SOTA shadow removal methods and three SOTA landmark detectors to study the robustness of shadow removal to facial landmark detection, and observe that there is a positive correlation between the shadow-removal accuracy and the landmark detection accuracy. The relationship is dominant especially when shadow degradation level is higher (*i.e.*, higher-severity shadow and adversarial shadow). It implies that feature embedding spaces of shadow removal and landmark detection aiming to optimize partially overlap with each other, which provides a bridge for the two tasks. Further, we propose a new shadow-removal framework regularized by landmark detection to further improve the visual quality and landmark detection simultaneously.

1.4 Proposed Approaches

As mentioned in Section 1.3, we propose three approaches to explore the applications of image restoration under adverse illumination. Each approach is proposed in the scenario of a practical image restoration problem, corresponding to shadow removal for image quality enhancement, night-to-day image translation for nighttime vehicle detection improvement, and shadow removal for image quality enhancement and facial landmark detection improvement simultaneously, respectively. Here we briefly introduce three problems and the proposed approaches.

Shadow Removal for Image Visual Quality Enhancement via Auto-exposure Fusion

In the first work, we explore the application of CNN-based shadow removal for image quality enhancement. Shadow removal is still a challenging task due to its inherent

background-dependent and spatial-variant properties, leading to unknown and diverse shadow patterns. Even powerful deep neural networks could hardly recover traceless shadow-removed background. With this motivation, we formulate shadow removal as an exposure fusion problem to address the challenges. We first estimate multiple over-exposure images with respective to the input shadow image to let the shadow regions in these images have the same color with shadow-free areas in the input image. We propose the shadow-aware FusionNet to generate fusion weight maps for fusing multiple over-exposure images and the shadow image to generate the final shadow-free counterpart. Moreover, we propose the boundary-aware RefineNet to eliminate the remaining shadow trace further. We conduct extensive experiments on the ISTD [117], ISTD+ [69], and SRD [96] datasets and achieve better performance in shadow regions over the state-of-the-art methods.

Detail-preserving Night-to-day Image Translation for Nighttime Vehicle Detection Improvement

In the second work, we explore the application of CNN-based night-to-day translation for vehicle detection improvement. Currently, object detection shows remarkable efficiency and reliability in standard scenarios such as daytime scenes with favorable illumination conditions. However, in face of adverse conditions such as the nighttime, object detection loses its accuracy significantly. We propose a framework to alleviate the accuracy decline when object detection is taken to adverse light conditions by using image translation method. We propose to utilize style translation based StyleMix method to acquire pairs of daytime image and nighttime image as training data for following nighttime to daytime image translation. To alleviate the detail corruptions caused by Generative Adversarial Networks (GANs), we propose to utilize Kernel Prediction Network (KPN) based method to refine the nighttime to daytime image translation. Experiments on vehicle detection demonstrate that our proposed method

achieves effective and accurate nighttime detection results.

Shadow Removal for Image Visual Quality and Facial Landmark Detection Improvement Simultaneously

In the last work, we extend the application from image quality enhancement/object detection improvement to consider both. Facial landmark detection is a very fundamental step of many downstream face-related vision applications. In practice, the facial landmark detection can be affected by a lot of natural degradations, where shadow is one of the most common and important type. Many advanced shadow removal methods have achieved impressive image quality, however, their effects on facial landmark detection are not well explored. For example, it remains unclear whether the shadow removal could enhance the robustness of facial landmark detection to diverse shadow patterns or not. In this work, for the first time, we construct a novel benchmark to link the two independent but relatable tasks (*i.e.*, shadow removal and facial landmark detection). In particular, the proposed benchmark SHAREL covers diverse face shadows with different intensities, sizes, shapes, and locations. Moreover, to mine hard shadow patterns against facial landmark detection, we propose a novel adversarial shadow attack, which allows us to construct a challenging subset of the benchmark for a comprehensive analysis. With the constructed benchmark, we conduct extensive analysis on three state-of-the-art shadow removal methods and three landmark detectors. The analysis of this work motivates us to design a novel landmark detection-aware shadow removal framework, which achieves higher restoration quality and facial landmark detection performance as well.

1.5 Structure of the Dissertation

This dissertation is organized as follows. Chapter 2 introduces the deep neural networks for image restoration, object detection, and facial landmark detection as the

basis of this dissertation. Chapter 3 conducts a literature review for related works. Chapter 4 presents a novel shadow removal framework via auto-exposure fusion strategy for image quality enhancement. Chapter 5 describes a detail-preserving night-to-day image translation method for nighttime vehicle detection improvement. Chapter 6 introduces a novel shadow removal benchmark for image quality and facial landmark detection improvement simultaneously. The conclusion is presented in Chapter 7, which also discusses the future research.

Chapter 2

Background

In this chapter, we briefly review the background knowledge which will be helpful to better understand the three works discussed in the dissertation. In Sec. 2.1, we introduce the Convolutional Neural Networks (CNNs) based image restoration architecture, *i.e.*, U-Net, which is useful to understand the Chapter 4. Sec. 2.2 describes Generative Adversarial Networks (GANs) based image restoration architecture by performing image-to-image translation. We introduce CycleGAN [151] in detail. Next, we explain the deep learning based object detectors in Sec. 2.3, especially Faster R-CNN [98]. Sec. 2.2 and Sec. 2.3 cover the background need to know in Chapter 5. Sec. 2.4 provides the background for facial landmark detection task in Chapter 6, and we demonstrate SAN [22] in detail.

2.1 CNN-based Image Restoration

Image restoration aims to recover the original clean images given the degraded images as input. It is an ill-posed inverse problem, because there are infinite possible mappings between corrupted images and their corresponding clean versions. Due to rich learning representation ability, deep learning models have achieved state-of-the-art (SOTA) performance for image restoration by learning strong priors from massive training data. Deep learning neural networks (DNNs)-based image restoration methods, with diverse network architectures and designs, include Convolution Neural Networks (CNNs) and Generative Adversarial Networks (GANs).

Existing CNNs for image restoration mainly cover two architecture designs: 1) encoder-decoder and 2) high-resolution feature processing with single scale. The encoder-decoder design consists of two stages. The encoder stage progressively maps an input image to a low-resolution feature representation in the latent space, while the decoder phase applies a gradual reverse mapping to that feature map to obtain a predicted clean image with the original resolution. A representative network architecture of the encoder-decoder is U-Net, as shown in Fig. 2.1. The contracting

path (encoder) of U-Net is a typical CNN feature extraction architecture, which consists of a stack of convolution, rectified linear unit (ReLU), and max pooling layers. The expansive path (decoder) includes up-convolution layers to recover the resolution and concatenations with features from the contracting path to integrate multi-level information. The goal of the encoder is to learn the content information, while the decoder is to recover the appearance details. Even though encoder-decoder can learn a comprehensive context by spatial-resolution reduction, the down-sampling operations in the encoder stage may harm the spatial details, which are hard to recover in the decoder stage. The high-resolution networks can preserve more accurate details by employing single-scale feature maps, but they cannot effectively encode rich contextual information due to limited receptive field.

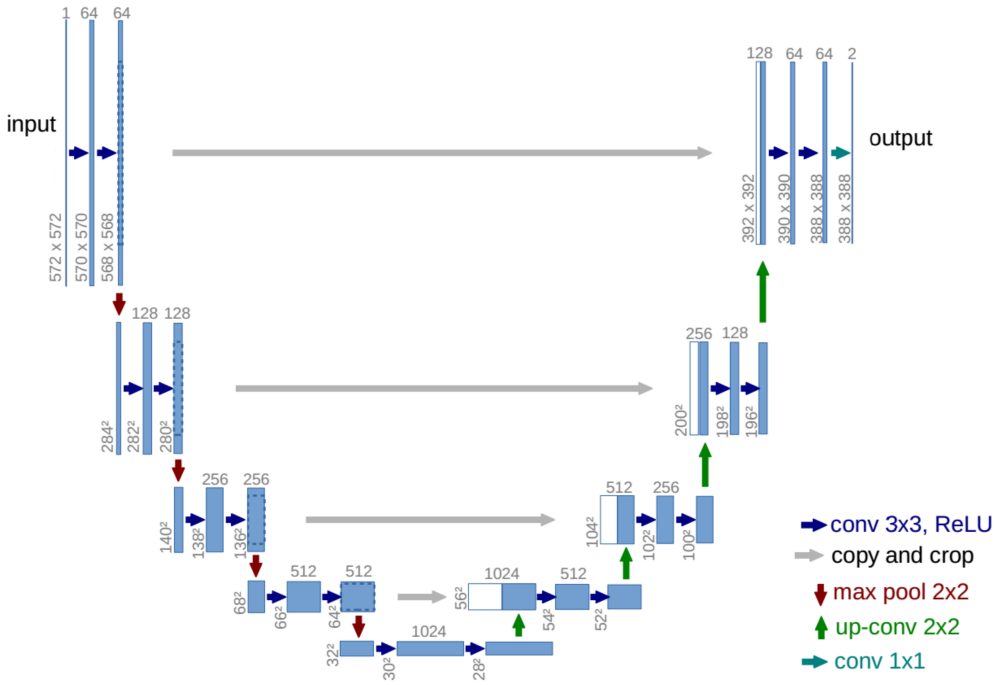

Figure 2.1: The network architecture of U-Net [99].

2.2 GAN-based Image Restoration

Encoder-decoder based image restoration methods are generally trained in a supervised way, requiring paired degraded images and clean images. However, such kind of data are not always available in the real-world applications. GAN-based image-to-image translation methods provide an effective way to restore clean images from degraded images in an unsupervised way with unpaired data. Degraded images can act as source domain data, while clean images are viewed as target domain data. GAN-based methods can perform image-to-image translation to reduce the domain distribution shift caused by image degradation between the source domain and the target domain. In general, GAN architecture includes a generator and a discriminator. The former is to generate synthetic images from the source input data, while the latter is to distinguish synthetic images from the real target data. The generator and discriminator compete with each other until the network is converged. As an extension of GAN, CycleGAN [151] can achieve cycle-level image-to-image translation between the source domain and the target domain, which is widely used for image restoration.

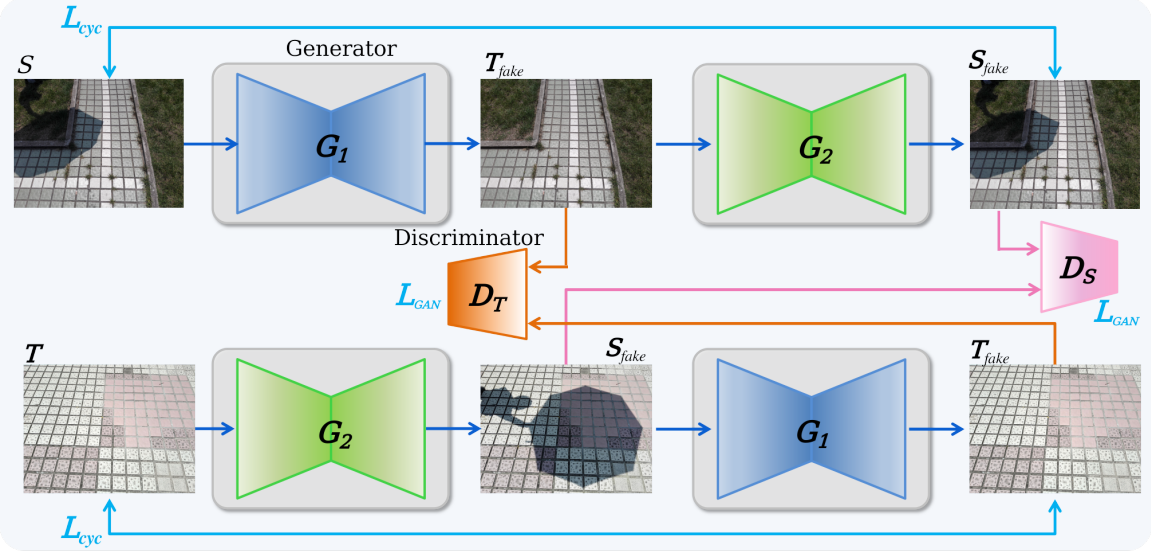

Figure 2.2: The network architecture of CycleGAN [151].

The image-to-image transfer in CycleGAN is achieved by training two generators and two adversarial discriminators, as shown in Fig. 2.2. We define the source domain (degraded images) as S and the target domain (clean images) as T. With regard to the translation between source domain S and target domain T, we define transfer functions G_1 and G_2 as the generators from S to T and from T to S, respectively. Meanwhile, two adversarial discriminators D_T and D_S, corresponding to G_1 and G_2, are defined. Specifically, feeding an image from the domain S to G_1 which acts like a Fully Convolutional Network (FCN) can generate a new image domain T_{fake}, and the discriminator D_T is to classify whether the new image in the domain T_{fake} is a real image from the domain T or a fake image generated by the generator G_1. Similarly, D_S aims to recognize whether an image is a real image from domain S or a fake image generated by G_2. Following [151], the total loss function of the domain adaptation is defined as

$$\begin{aligned} L(G_1, G_2, D_S, D_T, S, T) &= L_{GAN}(G_1, D_T, S, T) \\ &+ L_{GAN}(G_2, D_S, T, S) + \lambda L_{Cyc}(G_1, G_2, S, T), \end{aligned} \tag{2.1}$$

where λ is a weight to balance the adversarial training loss L_{GAN} and the cycle consistency loss L_{cyc} in the cycle architecture. L_{cyc} is to keep the transfers from S to T and from T to S cycle-consistent, which is defined as

$$L_{cyc}(G_1, G_2, S, T) = E_{i_S \sim I_S}[\| G_2(G_1(i_S)) - i_S \|_1] + E_{i_T \sim I_T}[\| G_1(G_2(i_T)) - i_T \|_1], \tag{2.2}$$

where $i_S \in I_S$ and $i_T \in I_T$ represent any images in S and T, respectively. The adversarial training loss function is defined as

$$L_{GAN}(G_1, D_T, S, T) = E_{i_T \sim I_T}[\log(D_T(i_T))] + E_{i_S \sim I_S}[\log(1 - D_T(G_1(i_S))]. \tag{2.3}$$

The training of these generators and discriminators aims to solve the optimization problem of

$$G_1^*, G_2^* = \arg \min_{G_1, G_2} \max_{D_S, D_T} L(G_1, G_2, D_S, D_T, S, T). \tag{2.4}$$

After solving Eq. (2.4) by gradient descent and back propagation, the learned generator G_1^* can be used to transfer the degraded images to the corresponding clean images.

2.3 Deep Learning Based Object Detection

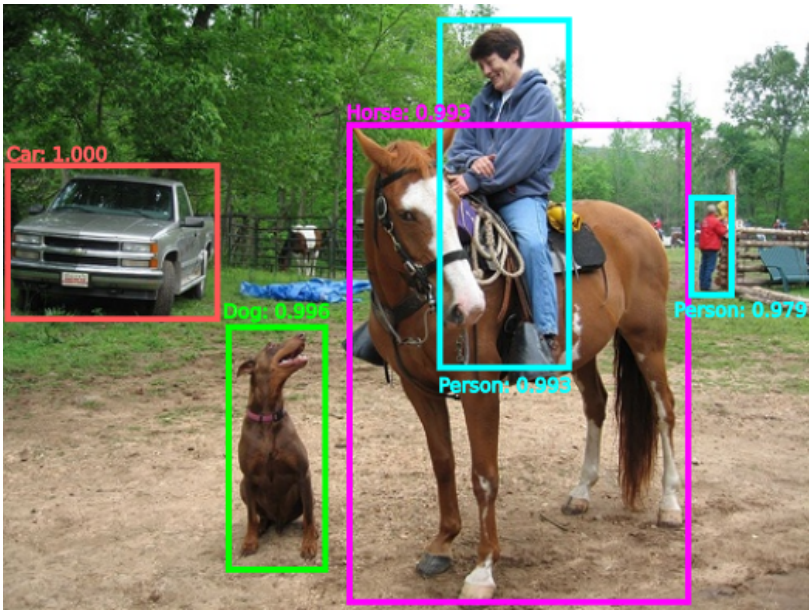

Figure 2.3: Illustration of object detection [98] task.

Object detection task aims to figure out not only the class labels of objects inside the input image, but also the localizations of the objects. Taking an image as the input, the object detector will output a classification probability and a bounding box for each object inside the image, as shown in Fig. 2.3. Current state-of-the-art general-purpose object detectors are CNN-based networks, including one-stage detectors, *e.g.*, single-shot multi-box detector (SSD) [86], and two-stage detectors, *e.g.*, Fast R-CNN [32] and Faster R-CNN [98]. Faster R-CNN is arguably the most popular two-stage object detector. Its first stage is to identify and generate region-of-interests (RoIs), and the second stage is to assign category probabilities to RoIs and refine their localizations. On the other hand, as a representative of the one-stage

detection frameworks, SSD can directly regress and classify anchors [86] on the whole image. While one-stage detectors tend to be more efficient without RoI training, two-stage detectors are often viewed to be more flexible and accurate. As an example, we briefly introduce the framework of Faster R-CNN.

The utility of CNNs for object detection was better established by a series of R-CNNs [33, 32, 98], where 'R' stands for regions or region proposals. R-CNN [33] performs detection on 2,000 region proposals, while a faster version of R-CNN, *i.e.*, Fast R-CNN [32], operates classification and regression directly on the whole image. Built on Fast R-CNN, Ren *et al.* [98] proposed a faster version in which a novel Region Proposal Network (RPN) shares the convolutional layers with Fast R-CNN to improve the computation efficiency.

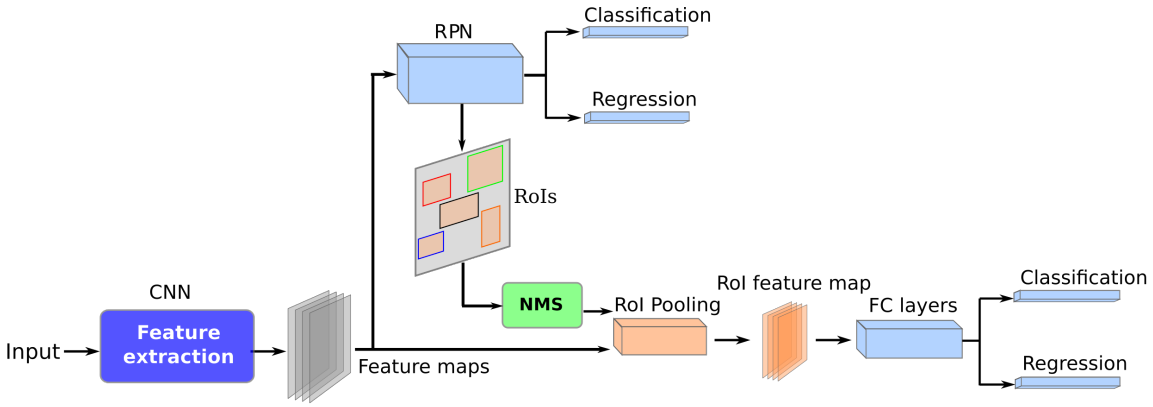

Figure 2.4: Illustration of Faster R-CNN [98] framework.

The pipeline of Faster R-CNN algorithm is shown in Fig. 2.4. Under Faster R-CNN, an input image is first fed to a backbone network, *e.g.*, VGG-16 [107], for feature extraction to generate feature maps. Following that, RPN aims to generate regions of interest (RoIs), *i.e.*, regions in the image that could potentially contain objects of interest, from a series of predefined anchors on the feature maps. Since predefined anchors are extracted in a pixel-wise way, most of RoIs are quite redundant due to overlapping with each other. Hence, RoIs will go through non-maximum suppression (NMS) [98] to filter out the redundant ones according to their classifica-

tion scores. RoIs with higher classification scores will be kept and fed into subsequent sub-networks. They will first go through an RoI pooling layer, which transforms all the RoIs' sizes to a predefined fixed one. Fixed-size feature maps (*i.e.*, RoI feature map) are then fed into two fully connected (FC) layers: one (Classification) predicts the category probabilities for each object, and the other (Regression) estimates the offset values of a target localization of an object corresponding to the RoI. In the inference phase, Faster R-CNN outputs a bounding box, the corresponding category label as well as the classification confidence for each object. The bounding box has four parameters $[x, y, w, h]$, where (x, y) denotes the center location, w and h are the width and height of the bounding box, respectively. We define the regression offsets $[t_x, t_y, t_w, t_h]$ from a RoI to the predicted bounding box as

$$\begin{aligned} t_x = (x - x^r)/w^r, t_y = (y - y^r)/h^r, \\ t_w = \log(w/w^r), \quad t_h = \log(h/h^r), \end{aligned} \tag{2.5}$$

and their ground truth $[t_x^*, t_y^*, t_w^*, t_h^*]$, *i.e.*, regression offsets from a RoI to the target location of an object as

$$\begin{aligned} t_x^* = (x^* - x^r)/w^r, t_y^* = (y^* - y^r)/h^r, \\ t_w^* = \log(w^*/w^r), \quad t_h^* = \log(h^*/h^r), \end{aligned} \tag{2.6}$$

where $[x^r, y^r, w^r, h^r]$ are 2D center location, width and height of the RoI and $[x^*, y^*, w^*, h^*]$ are 2D center location, width and height of the ground-truth localization of the object. The regression loss is defined as the Smooth $L1$ loss function [33] between the offsets of the predicted bounding box to the RoI and their ground truth if there is an object inside the RoI. The classification loss is defined as the log loss over two classes (object vs. no object).

Precision, Recall, mean Average Precision (mAP) are widely used metrics [98] to evaluate object detection performance. The steps of determining mAP are: 1) Compute a precision/recall based on a prediction result of inference against its ground

truth. A prediction is viewed as a true positive when the predicted category is the same as that of ground truth and the intersection-over-union (IoU) between the predicted bounding box and its ground truth is more than 0.5; 2) Calculate the AP as the area under the updated precision/recall curve by numerical integration [98]; 3) Average the AP among all classes as the mAP.

2.4 Facial Landmark Detection

The face, embodying rich nonverbal information, like identity and emotion, is of great importance in visual communication for the computer vision community. As a fundamental face-related task, facial landmark detection aims to automatically extract the localization of facial key landmark points, as shown in Fig. 2.5. Facial landmark detection determines the face shape by identifying the localization of characteristic landmarks (points that delineate eyebrows, eyes, nose, mouth, and face contour). Some of those key points are dominant points presenting the specific location of a facial component (*e.g.*, eye corner and mouth corner). Others are interpolated points connecting such dominant points between the facial components and the facial contour. Taken a facial image as input, the output of the facial landmark detector is a vector of 2D landmark coordinates.

Facial landmark detection is a fundamental step of many face-related tasks, *e.g.*, face recognition [153, 87], facial expression recognition [102], and head pose estimation [127]. They heavily rely on accurate detection of those landmark points. In specific, facial landmark points can provide salient features as a guidance for facial expression recognition to target at relevant regions of the face. Face alignment, *i.e.*, facial landmark detection, is a pre-processing step for face recognition task, which performs registration and alignment for facial images to eliminate in-plane rotations and provides facial crops for following processing. For head pose estimation, facial landmark points are critical to accurately estimate the parameters of 3D Morphable

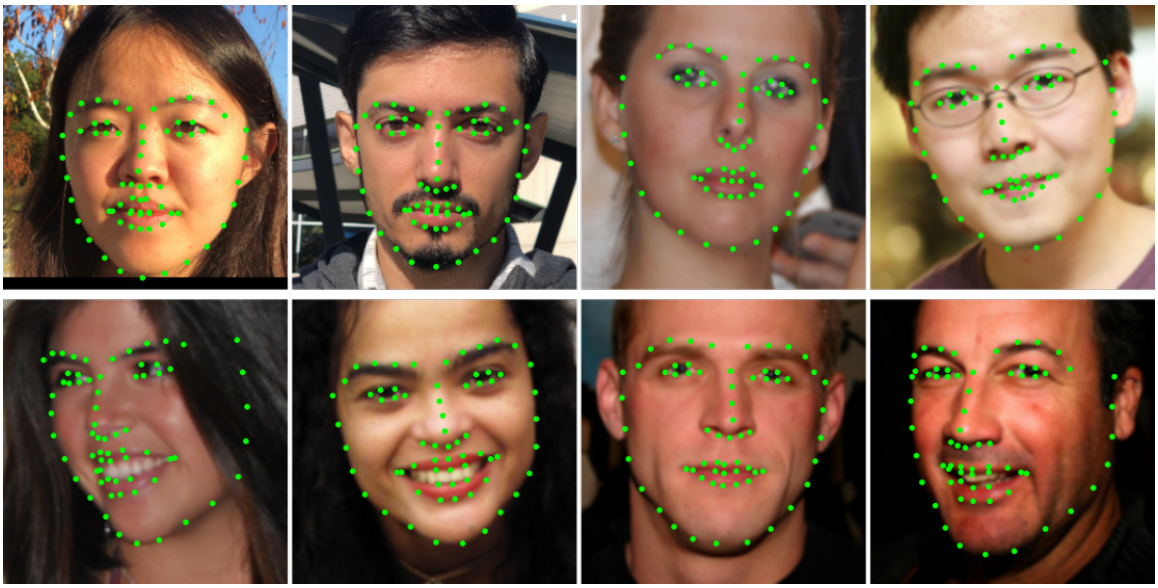

Figure 2.5: Examples of facial landmark detection.

Model (3DMM) [5]. Such parameters can construct the 3D head pose and facial action units. Facial landmark detection is very common in many real-world applications, including human-machine interaction, entertainment, security surveillance, and augmented reality.

There are several challenges for facial landmark detection: ① facial appearances pose significant variation across subjects under diverse facial identities, expressions, and head poses. ② The environment conditions, *e.g.*, adverse illumination and occlusion by external objects, will affect the facial appearance and degrade facial image quality. ③ Facial occlusion, by itself due to facial geometry or other objects, will lead to incomplete facial appearance.

In general, facial landmark detection algorithms first extract facial features, such as appearance and shape, then map them into landmark coordinates or heatmaps. The facial appearance represents the patterns of distinctive pixel intensity presented around the facial key points or in the whole face region, while face shape patterns are determined by the landmark locations and their spatial relationships. Traditional facial landmark detection models [14, 15, 16, 104] usually utilize the parametric models

to increase the shape variation. In specific, Active Shape Model (ASM) [15] proposes to model landmark distribution via Principal Component Analysis (PCA), and adjusts parameters to perform model matching with face images. However, traditional facial landmark detectors fail in the cases with cluttered background, severe occlusions, and higher variations on facial poses.

Recently, due to rich feature representation ability, deep learning based facial landmark detectors have achieved state-of-the-art detection performance, which include coordinate regression models and heatmap regression models. The former aims to directly map the facial image to the landmark coordinates with multiple regression models. For example, LAB [126] utilizes stacked hourglass network to extract more effective facial boundary features by employing the adversarial strategy and using message passing layers. LAB enhances the shape variations and improves the alignment performance. Nevertheless, regression of landmark coordinates mainly utilizes full connection operations, which lacks of fully making use of the spatial context embedding in the spatial space. Heatmap regression-based methods estimate landmarks by landmark heatmap regression, which better encode the context information and achieve state-of-the-art performance. Style Aggregated Network (SAN) [22] proposes a heatmap regression-based alignment method under image style variation. LUVLi [66] presents a novel end-to-end framework by jointly estimating the landmark locations, uncertainty, and visibility, achieving better detection performance.

We take SAN as an example to describe the face alignment algorithm in detail. The network framework of SAN is motivated by the widely-existing style variation in facial landmark detection benchmarks [101, 126]. Because the facial images are captured in the wild with unconstrained environment and various camera settings, *e.g.*, dark or light, gray-scale or color images, style variance is relatively large in the real-world facial landmark detection data. To solve this issue, SAN first transforms images with diverse styles into an aggregated style by employing CycleGAN, and

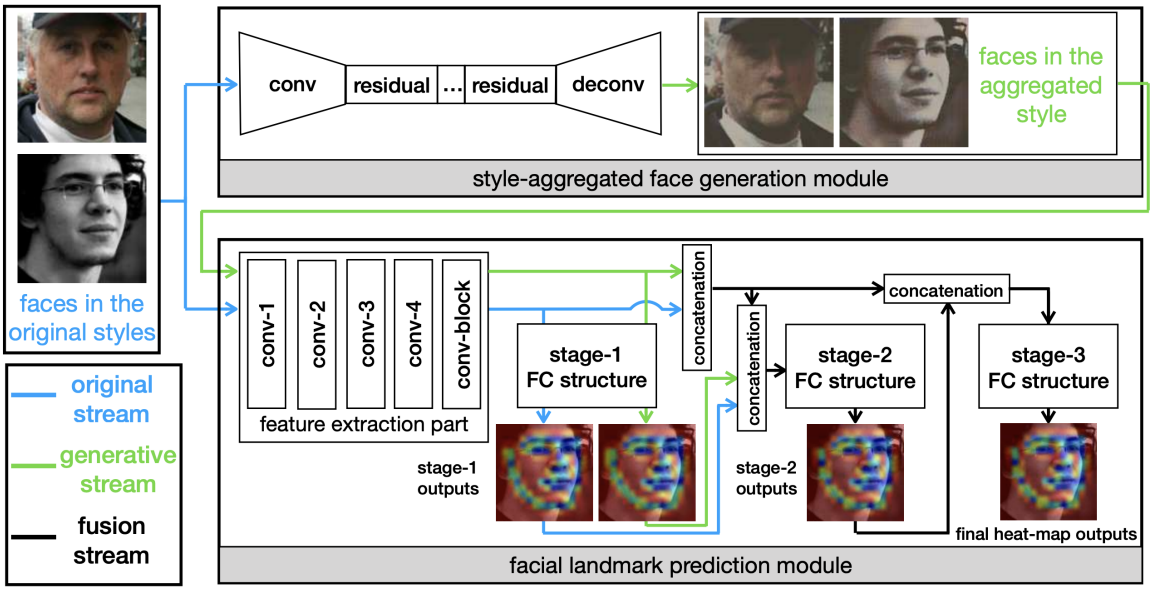

Figure 2.6: The framework of SAN [22].

then the original image and the style-aggregated image are fed into a deep model together to estimate the landmarks. In this way, the original image can provide localization information for landmark estimation, which can not be achieved by the style-aggregated image since it lacks fine details. The landmark detection pipeline of SAN is shown in Fig. 2.6. For the facial landmark prediction module, the green line refers to the style-aggregated faces and the blue line indicates the original face images. Both two kinds of images go through several convolution layers for feature extraction to obtain two complementary features. Such complementary features are fused to generate heatmap estimations by fully connected layers in a cascaded manner. Then the coordinate of each landmark is generated by performing the argmax function on each heatmap. SAN is optimized by the distance between the predicted heatmaps and the corresponding ground truth.

Chapter 3

Literature review

3.1 Image Restoration

Image restoration has been a long-standing topic in digital image processing and remains active in recent years with the rapid rise of deep learning techniques. Image restoration aims to recover a clean image from a degraded version. It is a typical ill-posed inverse problem, since there are infinite possible mappings between multi-dimensional degraded images and the restored ones. Traditional image restoration methods generally employ mathematics and probabilistic algorithms with handcrafted features to solve inverse problems, while they suffer from limited generalization ability. Recently, deep learning neural networks (DNNs) have achieved great success in computer vision tasks, such as recognition, classification, and image restoration. It is efficient for DNNs to learn image priors as effective regularization terms to restore degraded images. CNN-based image restoration methods learn mappings between degraded and clean images from large-scale paired datasets. For example, SRCNN [21] presents an image super resolution method, DnCNN [141] proposes a CNN-based image denoising method, AEFNet [26] is proposed for single image shadow removal, and Zero-DCE [36] is a low-light image enhancement network. Numerous CNN-based models employ larger and deeper neural network architecture designs to improve feature representation ability, such as residual block [64, 140], dense block [148], and others [68, 119, 109, 142, 125, 44]. Attention mechanism, such as channel attention [147, 19] and non-local attention [81, 89], is incorporated into CNN framework to improve the performance. In addition, generative adversarial networks (GANs) can also be applied to image restoration tasks [53, 67, 74]. GAN-based methods have achieved advanced performance in various applications including underwater imaging [72, 71], shadow removal [53], fluorescence image reconstruction [4], and image deblurring [67].

3.2 Shadow Removal

Shadows are present in most natural images where the light source is blocked. Spatial-variant color and illumination distortion presented in the shadow region can hinder the performance of other computer vision tasks [17, 62, 92, 103, 144], such as object detection and tracking, object recognition, semantic segmentation, *etc.*

For shadow images, many image restoration algorithms have been developed for removing the shadow and restoring the original clear image. Traditional shadow removal methods employ prior information, *e.g.*, gradient [35], illumination [143, 106, 129], and region [42, 116], for removing shadows. Recent deep learning based shadow removal methods boost the removal performance because of the available large-scale datasets of paired and unpaired shadow and shadow free images [69, 20, 53]. The Deshadow-Net [96] extracted multi-context features, involving global localization, appearance, and semantics, to predict a shadow matte layer for removing shadow in an end-to-end manner. Wang *et al.* proposed ST-CGAN [117] for joint shadow detection and removal by employing a stacked conditional GAN framework. The DSC [152] additionally utilized direction-aware context to improve shadow detection and removal. Le *et al.* [69] proposed to remove shadows from the perspective of shadow decomposition. On the other hand, the GAN based methods, *e.g.*, MaskShadow-GAN [53], made it possible to perform shadow removal on unpaired shadow and shadow free images by viewing it as an image-to-image translation problem. However, these methods suffered from artifacts and image blur. They also required the unpaired shadow and shadow free image sets to have similar statistical distribution.

In this dissertation, we model the shadow removal problem from a novel direction, *i.e.*, an auto-exposure fusion problem on paired shadow and shadow free images. Multiple over-exposure shadow images are generated to compensate the color and illumination degradation in the shadow region, then they are 'smartly' fused together to obtain the shadow free image.

3.3 Object Detection at Nighttime

The state-of-the-art performance in object detection is rapidly improving in recent years. One-stage (SSD [86], YOLO [97], RetinaNet [80]) and two-stage (Faster R-CNN [98], Mask R-CNN [48]) detection frameworks have achieved promising performance in real-world applications. We briefly review the underlying pipeline for Faster R-CNN algorithm. Under Faster R-CNN, an input image is first fed to a backbone network, e.g., VGG-16 [107], for feature extraction to generate a feature map. Then, feature maps are fed into two sub-networks: RPN phase and R-CNN (Region-CNN) phase. The RPN phase is mainly learning to generate Region of Interests (RoIs) for objects, and the R-CNN phase is to perform the object classification and refine the localization of RoIs. The output of Faster R-CNN is a bounding box for each object and corresponding class category and confidence. The bounding box has four parameters $[x, y, w, h]$, where (x, y) denotes the center location, w and h are the width and height.

Deep object detectors generally require a large amount of manually labeled data for supervised learning. Nonetheless, most of them operate well at daytime, under favorable illumination conditions, and scale badly to nighttime scenarios with challenging lighting conditions. Further, manual annotation of nighttime images are hard and time-consuming, because even human cannot clearly discern objects in adverse nighttime scenario. Nighttime detection task has attracted a lot of attention recently. Domain specific works [131, 31, 13] explore the human detection at nighttime by considering the type of cameras. Other works [65, 105] pertain to vehicle detection in driving scenarios. Domain-invariant representations [1, 100] or fusion works [114] are designed to be robust to illumination changes. Image translation work [2] aims to improve retrieval-based localization at nighttime.

In this dissertation, we focus on vehicle detection at nighttime in traffic surveillance scenarios. One traditional way to solve this is to fine-tune the already-trained

daytime perception model on the limited nighttime data, and hopefully it can perform well on nighttime scenarios, but it requires extra time and additional-labeled nighttime data for model fine-tuning. Another traditional way [73] might use Generative Adversarial Networks (GANs) based image to image translation methods in an unpaired way, such as CycleGAN [151] and UNIT [84], to transfer daytime images to fake nighttime images. Paired daytime and nighttime images are hard to obtain in the real-world applications, due to the dynamic traffic and environment changes. This kind of image translation considers this problem as domain adaptation for model fine-tuning on synthetic nighttime images without labeling the nighttime data. However, these methods also need extra time for model fine-tuning. In addition, GAN based image translation suffers from model collapse and does not preserve content details very well [93, 84, 151, 54, 121]. Bottleneck layers in a general deep generator hurt the learning ability of convolution kernels due to downsampling and upsampling operations, resulting in possible losing some structure details. Besides, unpaired training data of different domains limits the detail-preserving ability of generators due to the lack of pixel-wise correspondence. In contrast, we aim to adapt the daytime detection model to nighttime detection for reusing the daytime domain knowledge.

3.4 Shadow Degradation and Facial Landmark Detection

Facial landmark detection [128, 149, 61] is a fundamental step for numerous facial related applications, *e.g.*, face recognition and verification [153, 87], 3D face reconstruction [83], and safety-critical applications, *e.g.*, deepfake detection [150, 79], and facial reenactment [139, 110] for virtual avatar applications. It aims to detect the location of predefined facial landmarks, *e.g.*, the corners of the eyes, eyebrows, the tip of the nose. Deep facial landmark detectors can be classified into two types: direct coordinate regression [113, 115, 76] and heatmap-based approaches [22, 118, 154]. Coordinate-based landmark detection attempts to locate landmarks directly from images. Valle

et al. [115] infer landmark locations by a combined network with a tree-structure regression. Heatmap-based methods estimate a likelihood heatmap for each landmark and then infer localization prediction, rendering promising performance over direct regression [3]. Dong *et al.* [22] propose a style-aggregated network (SAN) to reduce the effect of style variations. Wang *et al.* [118] propose High-Resolution Network (HRNet) to fully explore high resolution information via performing multi-resolution fusion. LUVLi [66] proposes a deep model to jointly estimate the landmark locations and uncertainty predictions. Graph-based deep learning can also be utilized for facial landmark detection with good robustness and accuracy [76].

While recent deep-learning techniques bring us continuously improved landmark-detection performance, most of them are designed to handle only images of "clean faces". However, in real-world applications, face images usually contain image degradations, such as noise, shadow, and haze, which may significantly affect the performance of landmark detectors. Particularly, as a natural phenomenon, shadows are very common on face images – in practice, light to any face region can be occluded by surrounding objects, especially for portrait images captured in the wild. Spatial-variant illumination and color distortion in the shadow region [26] degrade the image quality and undermine the image features significantly. Shadow degrades visual quality, resulting in data distribution shift from clean images. Generally, domain gap will lead to performance drop when a pre-trained deep model on clean images is evaluated on degraded domain [108, 25]. The effect of shadow on facial landmark detection task is still under-explored. Image-level degradation via shadow can be alleviated by shadow removal such that providing a high-quality image for better visual effect. However, visual quality improvement does not always promise performance increasing of a high-level perception task [94, 50, 25]. Whether shadow removal benefits facial landmark detection remains unexplored. In this dissertation, we firstly attempt to explore the mutual influence of shadow removal and facial land-

mark detection to verify the effect of shadow removal on image visual quality and facial landmark detection.

Chapter 4

Shadow removal for image visual quality enhancement via auto-exposure fusion

4.1 Overview

Shadows are present in most natural images where the light source is blocked. Spatial-variant color and illumination distortion presented in the shadow region can hinder the performance of other computer vision tasks [17, 62, 92, 103, 144], such as object detection and tracking, object recognition, semantic segmentation, *etc.*

Previous shadow removal works either model this task based on physical shadow models for paired shadow and shadow-free images [69] or model it as an image-to-image translation problem based on the generative adversarial networks (GAN) for unpaired shadow and shadow-free images [53]. However, the learned shadow removal transformations by GAN-based methods, *e.g.*, MaskShadowGAN [53], tend to generate artifacts and image blur. They also suffer from data distribution requirements, where they expect the unpaired shadow and shadow-free image sets to share statistical similarity [78], which is hard to be satisfied when data acquisition is unstable. On the other hand, the publicly available large-scale datasets of paired shadow and shadow-free images, such as SRD [96], ISTD [117], and ISTD+ [69], allow shadow removal tasks to learn a physically plausible transformation in a supervised way. In this chapter, we focus on paired training data to perform the shadow removal task.

Shadow casting decreases the image quality with color and illumination degradation, over-exposure of the shadow image is an effective way to enhance the image quality. Intuitively, fusing the over-exposed one and the original shadow image could obtain the desired shadow-free image. Recent shadow decomposition works [69, 70], based on physical shadow models, mainly learn to relight the shadow image to a lit version and then fuse them together to acquire the desired shadow-free image via a shadow matte. However, since shadow casting degrades the color and illumination across the spatial region in a background-dependent and spatial-variant manner (*i.e.*, the contiguous shadow cast on the background image may cause the shadow region to appear differently based on how the original shadow-free background region looks

like, as well as where the shadow is cast spatially on the background image), we argue that multiple over-exposure fusion allows much higher level of flexibility and can provide a better solution to compensate the shadow region to have the same color and illumination with its non-shadow area, and better recovers the underlying content of the shadow region.

Shadow removal is still a challenging task for powerful state-of-the-art deep neural networks (DNN). Unknown and diverse shadow patterns pose two challenges to existing DNN based solutions: ❶ Shadow removal is a background-dependent task, which requires DNN to not only recover the illumination and color consistency with the shadow free area but also to preserve the content underlying the shadow. The spatial-variant property of shadow area requires that the fusion should be 'smart' enough to adaptively select the desired over-exposure pixels from various images to obtain the final shadow-free version. ❷ It is hard to obtain traceless background due to inconsistent shadow patterns along the boundary and inside the shadow region.

In this chapter, we propose a novel method, named auto-exposure fusion network, for single image shadow removal, as shown in Fig. 4.1(A). We first utilize exposure estimation to learn multiple over-exposure images by compensating the shadow region with different exposure levels. Then we propose the *shadow-aware FusionNet* to produce fusion weight maps across all the over-exposure images for addressing the first challenge. It can 'smartly' select which over-exposed pixel is the best one to recover the position-specific background. The proposed method fuses the input image and its over-exposure versions in a pixel-wise way. Further, we propose a *boundary-aware RefineNet* to remove the remaining shadow trace for refining the removal result obtained in the previous step. Figure 4.1(B) shows that the proposed method can obtain traceless background image than the state-of-the-art methods SP+M-Net [69] and DSC [52]. The contributions of this research are:

- To the best of our knowledge, this is the first work to study the shadow removal

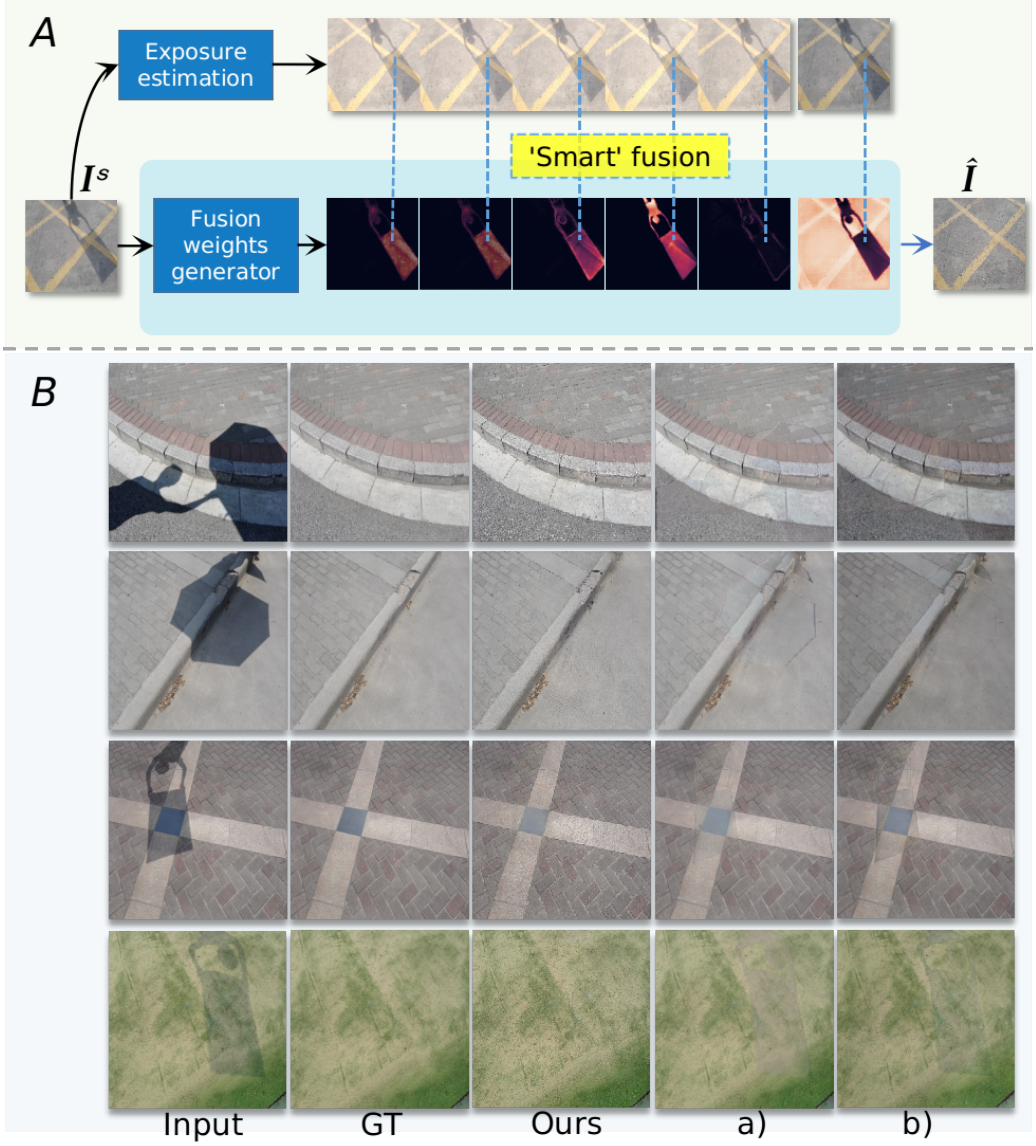

Figure 4.1: A: Illustration of the proposed auto-exposure fusion for shadow removal. B: Visualization results of our shadow removal results with the state-of-the-art methods. a) and b) are the shadow removal results of SP+M-Net [69] and DSC [52], respectively.

problem from the perspective of auto-exposure fusion.

- To accurately remove the shadow, we propose a new learning-based shadow-aware FusionNet followed by a boundary-aware RefineNet to accurately estimate, smartly fuse, and meticulously refine multiple over-exposure maps.
- The comprehensive experimental results on the public ISTD, ISTD+, and SRD

datasets show that the proposed method achieved better performance in shadow regions and comparable performance in non-shadow regions over the state-of-the-art methods.

4.2 Methodology

In this section, we propose to formulate the shadow removal as an exposure fusion problem to recover traceless background in the shadow image. We introduce the whole framework and reveal the challenges. Then, we explain how we generate the multi-exposure images for fusion. Next, our two main contributions, *i.e.*, *shadow-aware FusionNet* and *boundary-aware RefineNet*, help to address the challenges and achieve much better deshadowed images.

Exposure Fusion for Shadow Removal

We recast the shadow removal task as an exposure fusion problem and it can be formulated as

$$\hat{\mathbf{I}} = \phi(\mathbf{I}^{\mathrm{s}}), \tag{4.1}$$

where $\phi(\cdot)$ denotes a transfer function that can map the shadow image $\mathbf{I}^{\mathrm{s}}$ to the corresponding shadow free image $\hat{\mathbf{I}}$. A well-exposure image, *i.e.*, shadow free image, could be obtained by exposure fusion of brackets of multi-exposure images to improve the image quality of shadow image. The purpose of employing image over exposure is to compensate the shadow region to have the same color and illumination as the non-shadow region. In this research, we formulate the shadow region as an under exposed area of the shadow image. Then the problem left is to recover this area to its counterpart version which has the consistent color and illumination with the unshadowed area. Then, we can reformulate Eq. (4.1) to

$$\hat{\mathbf{I}} = \phi(\mathbf{I}^{\mathrm{s}}, \mathbf{I}_i^{\mathrm{o}}), i \in \{1, 2, \ldots, N\}, \tag{4.2}$$

where $\mathbf{I}_i^{\mathrm{o}}$ corresponds to the i-th over-exposure image of shadow image $\mathbf{I}^{\mathrm{s}}$. An intuitive way to solve it is to estimate an over-exposure version of the shadow image and then fuse them together to directly infer the desired shadow-free one. Nevertheless, shadow region is background-dependent and presents spatial-variant property, *i.e.*, the color and illumination distortion across shadow region is variant, single over-exposure could not adaptively reflect the degradation in spatial space.

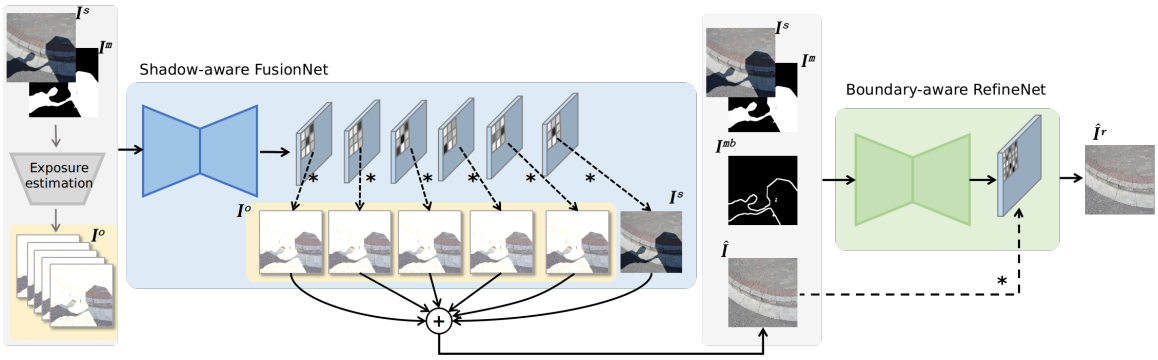

Figure 4.2: Illustration of the proposed framework for shadow removal with shadow-aware FusionNet and boundary-aware RefineNet.

Therefore, we propose an auto-exposure fusion network for fusing shadow image with a sequence of over exposed shadow images aiming to obtain the shadow free one. The whole framework of shadow removal is shown in Fig. 4.2. We first employ a deep learning network to generate a sequence of over exposed shadow images. Then we propose the *shadow-aware FusionNet* to 'smartly' fuse brackets of exposed images by generating fusion weight maps across each pixel of the input image to adaptively recover the color and illumination. However, due to the existing partial shadowed region, it is hard to obtain traceless background due to the inconsistent shadow patterns along the boundary and inside the shadow area. Further, we propose a *boundary-aware RefineNet*, to remove the residual shadow trace with the help of boundary mask.

Over-exposure Sequence Generation

We generate multiple exposure images through channel-wise weighting of the shadow image $\mathbf{I}^{\mathrm{s}}$ as following:

$$\mathbf{I}_i^{\mathrm{o}} = \alpha_i \mathbf{I}^{\mathrm{s}} + \beta_i, i \in \{1, 2, \ldots, N\}, \tag{4.3}$$

where $\alpha_i \in R^{3\times1}$ controls the exposure degree of the i-th over-exposure image and $\beta_i \in R^{3\times1}$ decides the potential intensity shifting. To realize the goal of shadow removal, we should estimate $\{\alpha_i\}$ and $\{\beta_i\}$ to make the shadow regions in the generated over-exposure images have the same color with the shadow-free regions in $\mathbf{I}^{\mathrm{s}}$. To this end, we aim to train a DNN to estimate the exposure parameters adaptively by taking the shadow image and shadow mask $\mathbf{I}^{\mathrm{m}}$ as input. Nevertheless, estimating all of the N exposure parameters directly via a DNN could let the training difficult. Instead, we adopt a two-stage way: first, we train a DNN to estimate the median exposure degree, *i.e.*, $\alpha_{\frac{N}{2}}$ and $\beta_{\frac{N}{2}}$

$$(\alpha_{\frac{N}{2}}, \beta_{\frac{N}{2}}) = \varphi(\mathbf{I}^{\mathrm{s}}, \mathbf{I}^{\mathrm{m}}), \tag{4.4}$$

where $\varphi(\cdot)$ denotes the DNN for exposure parameter estimation. Second, we generate all exposure images by performing a simple interpolation on $\alpha_{\frac{N}{2}}$ and $\beta_{\frac{N}{2}}$ with the assumption that the over-exposure sequence's images have similar color with minor difference

$$[\alpha_i, \beta_i] = \gamma_i[\alpha_{\frac{N}{2}}, \beta_{\frac{N}{2}}], i \in 1, 2, \ldots, N, \tag{4.5}$$

where $\{\gamma_i\}$ denotes the interpolation coefficients. Then, the key problem becomes how to train $\varphi(\cdot)$, which is a deep regression problem. The input data of exposure estimation is the shadow image and corresponding shadow mask. The ground truth of $\alpha_{\frac{N}{2}}$ $\beta_{\frac{N}{2}}$ is calculated by performing the least squares method [8] on the shadow mask covered regions of shadow image and its shadow-free counterpart. We optimize

the exposure estimation by minimizing the mean squared error (MSE) between the estimated parameters and its ground truth. Note that, exposure parameters are estimated independently between color channels to adaptively adjust color distortion caused by shadow as well as camera sensor.

Shadow-aware FusionNet

In this section, we design the FusionNet to fuse the generated over-exposure images $\{\mathbf{I}_i^{\text{o}}\}$ and produce the shadow-free image $\hat{\mathbf{I}}$. Intuitively, we can fuse $\{\mathbf{I}_i^{\text{o}}\}$ by assigning each pixel a weight across different exposure degree

$$\hat{\mathbf{I}}[p] = \sum_{i=0}^{N} \mathbf{W}_i[p]\mathbf{I}_i^{\text{o}}[p], \tag{4.6}$$

where $\mathbf{I}_0^{\text{o}} = \mathbf{I}^s$, and $\mathbf{W}_i$ has the same size with $\mathbf{I}_i^{\text{o}}$. Actually, such process means that each pixel of the final shadow-free image is the linear combination of N over-exposure images at the same pixel position and is fused independently. However, the fusion strategy ignores the local smoothness, leading to less natural or even noisy fusion results. Then, we further extend Eq. (4.6) by

$$\hat{\mathbf{I}}[p] = \sum_{i=0}^{N} (\mathbf{K}_i \circledast \mathbf{I}_i^{\text{o}})[p] = \sum_{i=0}^{N} \sum_{q \in \mathcal{N}(p)} \mathbf{k}_i^p[p-q]\mathbf{I}_i^{\text{o}}[q], \tag{4.7}$$

where $\circledast$ denotes the pixel-wise convolution, *i.e.*, each pixel is filtered by a kernel that is not shared by other pixels. Specifically, the p-th pixel of $\mathbf{I}_i^{\text{o}}$ (*e.g.*, $\mathbf{I}_i^{\text{o}}[p]$) and its neighboring pixels (*i.e.*, $\{\mathbf{I}_i^{\text{o}}[q] | q \in \mathcal{N}(p)\}$) are linearly combined by an exclusive kernel (*i.e.*, $\mathbf{k}_i^p$ the p-th kernel in $\mathbf{K}_i$) as the combination weights and $\mathbf{k}_i^p[p-q]$ denotes $[p-q]$-th elements of $\mathbf{k}_i^p$. $\mathcal{N}(p)$ is the neighboring pixels of p. Compared with Eq. (4.6), Eq. (4.7) considers the neighboring pixels' color and could avoid potential noisy results with better removal effect. We denote $\mathcal{K} = \{\mathbf{K}_i\}$ as pixel-wise fusion kernels.

Then, the key of generating the true shadow-free image is to estimate the fusion kernels accurately. Motivated by the above process, we propose to estimate the fusion

weight maps by training a CNN that takes the shadow image with shadow mask for guidance

$$\mathcal{K} = \text{FusionNet}(\mathbf{I}^{\text{m}}, \mathbf{I}^{\text{o}}), \tag{4.8}$$

where $\mathbf{I}^{\text{m}}$ is the shadow mask. The FusionNet is required to understand the shadow images and predict kernels that can spatially adapt to different shadow-covered contexts, thus can select suitable pixels from the multiple over-exposure images for shadow removal.

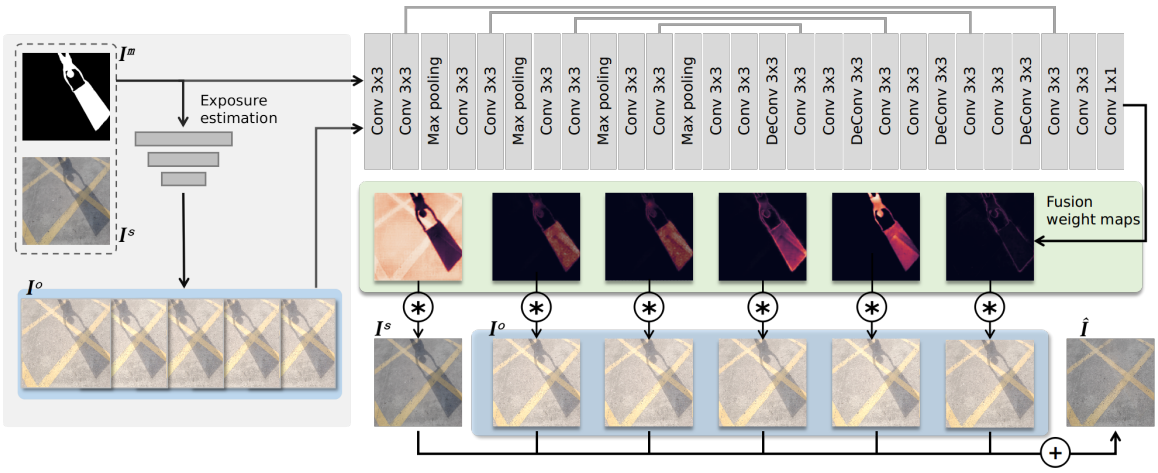

Figure 4.3: Illustration of the proposed shadow-aware FusionNet.

The pipeline of the shadow-aware FusionNet is shown in Fig. 4.3. FusionNet achieves shadow free recovery by 'smartly' selecting position-specific over-exposure pixels. The input data includes brackets of multiple exposure images, *i.e.*, the shadow image $\mathbf{I}^{\text{s}}$, corresponding shadow mask $\mathbf{I}^{\text{m}}$, and over-exposure images $\{\mathbf{I}_i^{\text{o}}\}$. FusionNet generates fusion weight maps, across all over-exposure images, to 'smartly' fuse the proper pixels from over-exposure versions with the shadow ones to the shadow free counterpart. Shadow mask $\mathbf{I}^{\text{m}}$ acts as a fusion guidance for FusionNet to let it assign low weights to non-shadow region and focus mostly on the shadow region, which is shown by the fusion weight maps in Fig. 4.3.

We employ L_1 distance to optimize our shadow-aware FusionNet. The loss function $\mathcal{L}_{\text{pix}}(\hat{\mathbf{I}}, \hat{\mathbf{I}}^*)$ is the pixel-wise L_1 distance between the ground truth shadow free image $\hat{\mathbf{I}}^*$ and the shadow removed image $\hat{\mathbf{I}}$

$$\mathcal{L}_{\text{pix}}(\hat{\mathbf{I}}, \hat{\mathbf{I}}^*) = \|\hat{\mathbf{I}}^* - \hat{\mathbf{I}}\|_1. \tag{4.9}$$

Boundary-aware RefineNet

Partially shadowed (penumbra) pixels exist along the shadow boundary. Inconsistent shadow patterns along the shadow boundary and inside the shadow region are still a challenge to state-of-the-art solutions to obtain traceless background. To solve this issue, we propose a boundary-aware RefineNet to eliminate the remaining shadow trace, which is shown in Fig. 4.4. It acts as a refinement of the shadow removal result obtained from FusionNet. Specifically, we model the boundary-aware RefineNet as

$$\mathcal{F} = \text{RefineNet}(\mathbf{I}^{\text{s}}, \mathbf{I}^{\text{m}}, \mathbf{I}^{\text{mb}}, \hat{\mathbf{I}}), \tag{4.10}$$

where $\mathbf{I}^{\text{mb}}$ is a penumbra mask, as shown in Fig. 4.4. Similar to Eq. (4.7), $\mathcal{F}$ is also pixel-wise refine kernels that integrate the context of pixel's $k \times k$ neighborhood region with that pixel to remove remaining trace. Then the refined shadow free image becomes

$$\hat{\mathbf{I}}^{\text{r}}[p] = (\mathbf{F} \circledast \hat{\mathbf{I}})[p] = \sum_{q \in \mathcal{N}(p)} \mathbf{f}^p[p-q]\hat{\mathbf{I}}[q] \tag{4.11}$$

where $\mathbf{f}^p \in \mathbb{R}^{k \times k}$ is the exclusive kernel for performing convolution between the $k \times k$ neighboring pixels of the pixel p (*i.e.*, $\mathcal{N}(p)$) and the kernel weights in $\mathbf{f}^p$.

RefineNet's input data includes: the shadow image $\mathbf{I}^{\text{s}}$, shadow mask $\mathbf{I}^{\text{m}}$, penumbra mask $\mathbf{I}^{\text{mb}}$, and initial shadow removal result $\hat{\mathbf{I}}$. Penumbra mask acts as a guidance for RefineNet to keep color and illumination consistency of the shadow removed, shadow boundary, and the non-shadow regions. Penumbra mask $\mathbf{I}^{\text{mb}}$ is extracted by computing the difference between dilated shadow mask $\mathbf{I}^{\text{md}}$ and eroded shadow

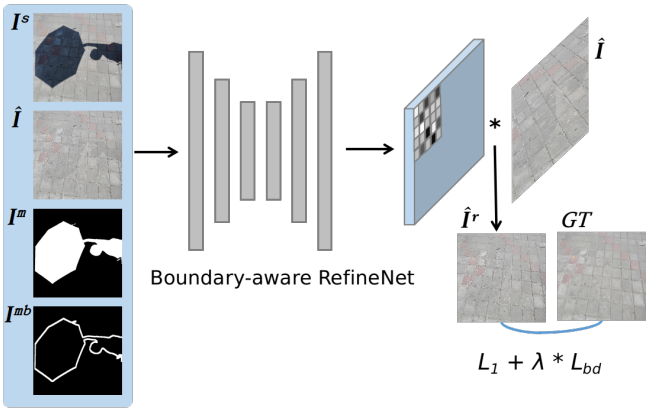

Figure 4.4: Illustration of the proposed boundary-aware RefineNet.

mask $\mathbf{I}^{\text{me}}$ for the penumbra region. We dilate/erode the shadow mask by 7 pixels to generate $\mathbf{I}^{\text{md}}$ and $\mathbf{I}^{\text{me}}$. The goal of RefineNet is to output a refined shadow removal image without trace.

The pixel-wise L_1 distance $\mathcal{L}_{\text{pix}}(\hat{\mathbf{I}}^{\text{r}}, \hat{\mathbf{I}}^{*})$, between the ground-truth shadow-free image $\hat{\mathbf{I}}^{*}$ and the refined version of shadow removed image $\hat{\mathbf{I}}^{\text{r}}$, is utilized to optimize the boundary-aware RefineNet. In addition, inspired by Poisson image editing [95], we propose a boundary-aware loss $\mathcal{L}_{\text{bd}}(\hat{\mathbf{I}}^{\text{r}}, \mathbf{I}^{\text{s}}, \hat{\mathbf{I}}^{*}, \mathbf{I}^{\text{m}})$ to seamlessly remove the shadow. It is defined as

$$\mathcal{L}_{\text{bd}}(\hat{\mathbf{I}}^{\text{r}}, \mathbf{I}^{\text{s}}, \hat{\mathbf{I}}^{*}, \mathbf{I}^{\text{m}}) = \text{MSE}(\nabla\hat{\mathbf{I}}^{\text{r}}, \nabla\mathbf{I}^{\text{s}}) * (1 - \mathbf{I}^{\text{m}}) + \text{MSE}(\nabla\hat{\mathbf{I}}^{\text{r}}, \nabla\hat{\mathbf{I}}^{*}) * \mathbf{I}^{\text{m}} \tag{4.12}$$

where ∇ denotes the Laplacian gradient operator. It aims to minimize the gradient domain along the shadow boundary. It keeps the same gradient domain of non-shadow region between predicted shadow-free image $\hat{\mathbf{I}}^{\text{r}}$ and shadow image $\mathbf{I}^{\text{s}}$. At the same time, it reduces the difference of gradient domain between predicted shadow-free image $\hat{\mathbf{I}}^{\text{r}}$ and ground-truth one $\hat{\mathbf{I}}^{*}$ in the shadow region. The total loss of RefineNet is a weighted sum of $\mathcal{L}_{\text{pix}}(\hat{\mathbf{I}}^{\text{r}}, \hat{\mathbf{I}}^{*})$ and $\mathcal{L}_{\text{bd}}(\hat{\mathbf{I}}^{\text{r}}, \mathbf{I}^{\text{s}}, \hat{\mathbf{I}}^{*}, \mathbf{I}^{\text{m}})$, as shown in Fig. 4.4. We set λ to 0.1 in the experiment.

Implementation Details

The proposed pipeline is implemented in PyTorch. The details of network setting and training are:

1) Exposure estimation is trained together with FusionNet. Its goal is to estimate the median-exposure version of the input shadow image. We employ ResNeXt [130] as backbone to do the estimation. We set the number of over-exposure images N to 5 by linearly interpolating the estimated exposure parameters with scaling factors between [0.95, 1.05]. For FusionNet, we employ a DNN with U-Net256 [99] as backbone.

2) Then boundary-aware RefineNet is to improve the shadow removal result with the same backbone as FusionNet. We train the RefineNet with exposure estimation and FusionNet together but freezing the latter two. Both FusionNet and RefineNet take the shadow mask as input, and we describe the setting of the datasets in Sec. 6.3.

In our experiments, same training parameters setting are employed for these three parts. The input image is resized to 256×256. The minibatch size is 8 and the initial learning rate is set to 0.0001. We use Adam optimizer for all the networks. We trained 400 epochs for each network.

4.3 Experiment

Datasets and Evaluation Measurement

Datasets. We train and evaluate the proposed method on three public datasets: ISTD [117], adjusted ISTD (ISTD+) [69], and SRD [96] datasets. They all have paired shadow and shadow-free images. Dataset ISTD and its adjusted version also have shadow masks. We introduce these three datasets as following:

1) The training set of ISTD dataset has 1,330 triplets of shadow, shadow free, and shadow mask images. The testing split consists of 540 triplets. The ISTD+

dataset has the same number of triplets with ISTD except that it adjusts the color inconsistency, between the shadow and shadow free image, with image processing algorithm [69]. The color mismatch results from the data acquisition setup. We use ground-truth shadow masks for training stage, while for inference, we compute the shadow masks by operating Otsu's algorithm to the difference between shadow and shadow free images, similar to MaskShadow-GAN [53]. We additionally refine these masks by a median filter to remove noises.

2) SRD dataset consists of 408 pairs of shadow and shadow free images without the ground-truth shadow mask. Here we simply use an adaptive threshold detection method, same as the one used in ISTD dataset, to extract the shadow mask from the difference between shadow free and shadow images. The extracted shadow masks are used both for training and testing. We utilize the public shadow masks provided by DHAN [18] for evaluation.

Evaluation measures. We utilize the root mean square error (RMSE) in LAB color space between the shadow removal result and the ground-truth shadow free image to evaluate the shadow removal performance, following previous works [117, 42, 96, 69, 18]. We directly compare our auto-exposure fusion framework with several state-of-the-art methods on the ISTD, ISTD+, and SRD datasets in quantitative and qualitative ways.

Shadow Removal Evaluation on ISTD Dataset

We first report the shadow removal results of our method on ISTD dataset [117], as shown in Table 4.1. We compare the proposed method with the state-of-the-art algorithms: Guo *et al.* [42], Gong *et al.* [34], ST-CGAN [117], MaskShadow-GAN [53], DSC [52], and DHAN [18]. Different from other methods, MaskShadow-GAN utilizes unpaired shadow and shadow free images for training. The first row shows the RMSE values of the input shadow image and corresponding shadow free image

Table 4.1: Shadow removal results of our networks compared to state-of-the-art shadow removal methods on the ISTD [117] dataset.

Method \ RMSE	**Shadow**	**Non-Shadow**	**All**
Input Image	32.12	7.19	10.97
Guo *et al.* [42]	18.95	7.46	9.30
Gong *et al.* [34]	14.98	7.29	8.53
MaskShadow-GAN [53]	12.67	6.68	7.41
ST-CGAN [117]	10.33	6.93	7.47
DSC [52]	9.76	6.14	6.67
DHAN [18]	8.14	6.04	6.37
Ours	**7.77**	**5.56**	**5.92**

without shadow removal operation. It shows that the proposed method obtains the best shadow removal performance in both shadow and non-shadow regions, leading to the lowest RMSE in the whole image. Specifically, the proposed method outperforms DSC [52] by 20.3% and 11.2% RMSE decreasing in shadow region and the whole image, respectively. The proposed method also outperforms the method DHAN [18] by reducing the RMSE from 8.14 to 7.77 in the shadow region. Training with unpaired data doesn't perform as well as training with paired version. Specifically, the proposed method outperforms MaskShadow-GAN by 38.6% and 20.1% RMSE decreasing in the shadow region and the whole area, respectively.

We also report the shadow removal performance of our proposed method on the adjusted ISTD (ISTD+) [69] dataset. As shown in Table 4.2, we compare the proposed method with state-of-the-art algorithms: Guo *et al.* [42], Gong *et al.* [34], ST-CGAN [117], DeshadowNet [96], MaskShadow-GAN [53], Param+M+D-Net [70], and SP+M-Net [69]. It turns out that the proposed method achieves the best shadow removal performance in the shadow region, outperforming SP+M-Net by 17.7% lower RMSE. It outperforms the DeshadowNet and ST-CGAN trained with paired shadow and shadow-free images, decreasing the RMSE by 59.1% and 51.4%, respectively. Compared to methods training with unpaired data, training with paired images

Table 4.2: Shadow removal results of our networks compared to state-of-the-art shadow removal methods on the ISTD+ [69] dataset.

Method \ RMSE	**Shadow**	**Non-Shadow**	**All**
Input Image	40.2	2.6	8.5
Guo *et al.* [42]	22.0	3.1	6.1
Gong *et al.* [34]	13.3	-	-
ST-CGAN [117]	13.4	7.7	8.7
DeshadowNet [96]	15.9	6.0	7.6
MaskShadow-GAN [53]	12.4	4.0	5.3
Param+M+D-Net [70]	9.7	**3.0**	4.0
SP+M-Net [69]	7.9	3.1	**3.9**
Ours	**6.5**	3.8	4.2

still acquire better results. The proposed method outperforms Param+M+D-Net by about 32.9%, trained with unpaired shadow and shadow free patches. The proposed method achieves the comparable performance in the non-shadow and whole image region.

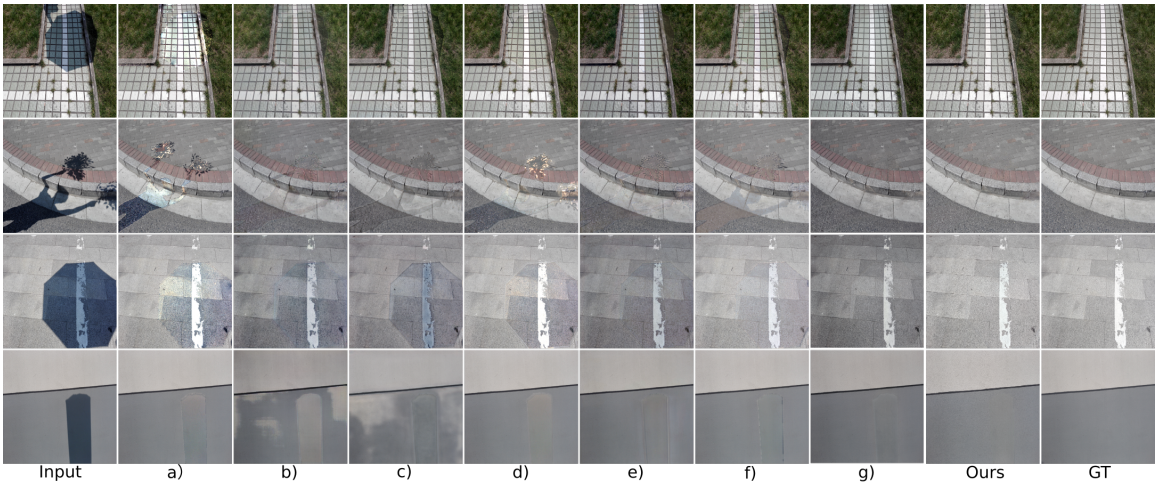

Figure 4.5: Illustration of the visualization results of shadow removal on dataset ISTD [117]. a) to g) are the results from comparison methods: Guo *et al.* [42], ST-CGAN [117], MaskShadow-GAN [53], Param+M+D-Net [70], DSC [52], SP+M-Net [69], and DHAN [18], respectively.

Figure 4.5 shows the visualization results of shadow removal from our methods and other state-of-the-art methods on the ISTD dataset. We can see that our result could

recover traceless background in the shadow region. We can clearly see that traditional method, Guo *et al.* [42], suffers from severe artifacts and could not recover shadowed pixels successfully due to limited feature representation ability. ST-CGAN could improve the performance by training large-scale data, while it tends to generate blurry images, random artifacts, and incorrect colors, *e.g.*, the fourth row shadow removed image. MaskShadowGAN and Param+M+D-Net also suffer from producing blurry images. Random artifacts along the shadow boundary can be easily spotted in the result of Param+M+D-Net, and it relights the boundary rather than removing it. Even though DSC and SP+M-Net could remove most of the shadow, their results still have trace along the shadow boundary, which does not exist in our result.

Shadow Removal Evaluation on SRD Dataset

In this section, we show our shadow removal results on SRD dataset [96] in Table 4.3. We evaluate our result with the public masks provided by DHAN [18]. The proposed method obtains the best shadow removal results with the lowest RMSE in the shadow region. It reduces the RMSE from 8.94 to 8.56, compared to DHAN.

Table 4.3: Shadow removal results of our networks compared to state-of-the-art shadow removal methods on the SRD [96] dataset.

Method \ RMSE	**Shadow**	**Non-Shadow**	**All**
Input Image	40.28	4.76	14.11
Guo *et al.* [42]	29.89	6.47	12.60
DeshadowNet [96]	11.78	4.84	6.64
DSC [52]	10.89	4.99	6.23
DHAN [18]	8.94	**4.80**	**5.67**
Ours	**8.56**	5.75	6.51

As shown in Table 4.3, the non-shadow region's RMSE values of different methods are very close (mean: 5.4, standard deviation: 0.6), which are similar to those of the Table 4.2 for ISTD+ dataset (mean: 4.4, standard deviation: 1.7). However, the

standard deviations of the RMSE values in shadow region are significantly larger. This means that different methods including ours all perform well and very close on the non-shadow region, and the main difficulty of this problem comes from the shadow region. For the shadow region, our method obviously obtains the best performance.

Ablation Study

We conduct ablation studies on ISTD+ dataset to evaluate the contribution of each step of our proposed method. For the effectiveness of per-pixel kernel fusion, *i.e.*, Eq. (4.7) over Eq. (4.6), we perform *Fusion-N1* which fuses pairs of over-exposure and shadow images with the per-pixel kernel that considers 3×3 neighboring pixels and with pixel-wise fusion. It turns out that fusing image pair with neighboring information can boost the performance from 7.6 to 7.1 for RMSE in the shadow region, because neighboring region provides important spatial context information to represent the structure. We set 3×3 neighborhood for FusionNet.

Then, we conduct experiments to verify the effectiveness of multiple over-exposure by controlling the number of over-exposure images. In our implementation, we set the number N to 1, 3, and 5. The shadow removal models are denoted as *Fusion-N1*, *Fusion-N3*, and *Fusion-N5*, respectively. N is set to 5 for the remaining experiments. We test the effectiveness of boundary-aware RefineNet and loss $\mathcal{L}_{\text{bd}}$ by models *Fusion+RefineNet* and *Fusion+RefineNet+*$\mathcal{L}_{\text{bd}}$, respectively. The results are summarized in Table 4.4.

To estimate the effectiveness of multiple over-exposure to the shadow-aware FusionNet, we report the performance in shadow, non-shadow, and whole image regions with the metric RMSE. When N is 5, the shadow removal result in the shadow region reaches lower RMSE 6.9, compared to when $N = 1$. We set N to 5 for later ablation experiments. With the introducing of boundary-aware RefineNet, *Fusion+RefineNet* improves the shadow removal performance by about 0.3 RMSE decreasing. It verifies

Table 4.4: Ablation study of shadow removal on the ISTD+ [69] dataset.

Method \ RMSE	**Shadow**	**Non-Shadow**	**All**
Input Image	40.2	2.6	8.5
Fusion-N1	7.1	3.9	4.4
Fusion-N3	7.2	3.9	4.5
Fusion-N5	6.9	4.0	4.4
Fusion+RefineNet	6.6	3.8	4.3
Fusion+RefineNet+$\mathcal{L}_{\text{bd}}$	**6.5**	**3.8**	**4.2**

that penumbra region is a challenge for shadow removal task to get traceless background. The RMSE in the non-shadow region also decreased, compared to *Fusion-N5*. Further, $\mathcal{L}_{\text{bd}}$ loss optimized the shadow removal model *Fusion+RefineNet+*$\mathcal{L}_{\text{bd}}$ better to reach the lowest RMSE 6.5, 3.8, and 4.2 in the shadow, non-shadow and the whole image regions.

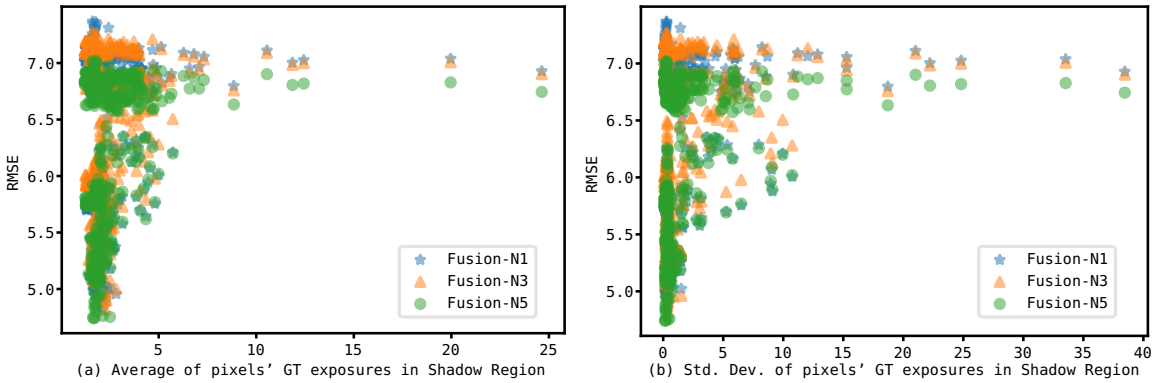

Figure 4.6: RMSE vs. average (*i.e.*, (a)) and std. dev. (*i.e.*, (b)) of pixels' GT exposures in shadow region for each testing example.

To explain the small margin of shadow removal performance gain of *Fusion-N5* over *Fusion-N1*, we calculate the ground truth exposure for the p-th pixel in the shadow region by dividing the shadow-free pixel with its shadow counterpart for each testing example. Then, we count the average and the standard deviation (std. dev.) of GT exposures of all pixels in the shadow region for each example and show their relationship to the example's RMSE of the 3 variants in Fig. 4.6. We see that: 1) For the most examples, *Fusion-N5* and *Fusion-N3* have lower RMSE than *Fusion-N1&N3*

and *Fusion-N1*, respectively. 2) Most examples' GT exposures have small variations (*i.e.*, small std. dev.) across spatial coordinates, leading to similar RMSE on the three methods. 3) When the GT exposures' variation become larger, the advantages of *Fusion-N3&5* become more obvious.

Table 4.5: Comparison of traceless background results in penumbra region on ISTD+ [69] dataset.

Method \ RMSE	**Penumbra**
SP+M-Net [69]	7.06
Ours	**5.96**

We also compare our method with the state-of-the-art method SP+M-Net [69] about measuring the shadow removal result without residual trace. We evaluate the RMSE metric in the penumbra region by considering the penumbra mask $\mathbf{I}^{\text{mb}}$. As shown in Table 4.5, our method performs better, decreasing RMSE by 15.5%. Visualizations are shown in Fig. 4.5(f) and ours. The SP+M-Net does not perform well to remove the residual trace.

Figure 4.7: Illustration of the visualization results of shadow removal for a) a fully shadowed image and b) a fully shadow-free image from ISTD+ [69] dataset.

We also show the visualization results of two edge cases sampled from ISTD+ [69] dataset for shadow removal, which are shown in Fig. 4.7. Figure 4.7(a) shows shadow removal result of a fully shadowed image. The input shadow mask has value 1 for all its elements, since the whole region needs to perform shadow removal. We also report the evaluation metric RMSE before and after shadow removal. Before shadow removal, the RMSE between the input shadow image and the ground-truth shadow-free image is 13.10, while it drops to 5.55 (*i.e.*, the RMSE between the shadow-removed image and the ground truth) after shadow removal. It shows the proposed method can effectively recover the image visual quality with lower RMSE. Figure 4.7(b) presents the shadow removal result of a fully shadow-free image. The corresponding shadow mask has value 0 for all its elements. Before shadow removal, the RMSE between the shadow image and the ground-truth shadow-free image is 1.80. After shadow removal, the RMSE between the shadow-removed image and ground truth increases to 2.50. The reason might be that the proposed method finally overexposes the non-shadow image after exposure fusion, which introduces undesired noise into the input image. This observation is consistent with the shadow removal results shown in Table 4.2 and Table 4.3, *i.e.*, the non-shadow region's RMSE values increase after shadow removal, while the proposed method achieves comparable performance with other methods in the non-shadow region.

4.4 Chapter Summary

In this chapter, we have proposed a novel and robust over-exposure fusion method for performing shadow removal task. Multiple over-exposure, relighting each pixel with different exposures, could compensate each pixel individually to tackle position specified color and illumination degradation. It benefits the shadow removal task by recovering the natural image from the spatial variant color and illumination degradation. Shadow-aware FusionNet smartly fuses brackets of over-exposure

shadow images with shadow image by an adaptive per-pixel kernel weight map. It helps to fully recover the background content preserving the color and illumination details. The proposed boundary-aware RefineNet further eliminates the remaining trace caused by the penumbra area along the shadow boundary. With the boundary loss added, by optimizing to preserve the non-shadow region and recover the ground-truth shadow-free area of the shadow image, our work can obtain traceless background with the state-of-the-art shadow removal performance on the ISTD, ISTD+, and SRD datasets.

Chapter 5

Detail-preserving night-to-day image translation for nighttime vehicle detection improvement

5.1 Overview

With the fast development of computer vision and deep Convolutional Neural Networks (CNNs), visual data understanding in image and video has attracted a lot of attention [146, 138, 132, 137, 45, 135, 75]. For example, in the Intelligent Transportation Systems (ITS), detecting the vehicles in each frame of the traffic surveillance video is important to extract the real-time traffic flow parameters [63] for the efficient traffic control and obtain the vehicle trajectories [10] for the calibrated traffic model simulation, *etc.*

Most of existing researches focus on daytime perception task through supervised learning, however, they generalize badly to adverse conditions such as nighttime scenarios [73]. The adversity of nighttime scenario poses two challenges for the success of perception task at nighttime: 1) nighttime data with a large amount of annotations is usually scarce compared to the large-scale daytime data, since accurate annotation of nighttime images is relatively hard to obtain. 2) The visual hazards, such as underexposure and noise, of nighttime images cause the extracted features corrupted.

One traditional way to solve this is to fine-tune the already-trained daytime perception model on the limited nighttime data, and hopefully it can perform well on nighttime scenarios, but it requires extra time and additional-labeled nighttime data for model fine-tuning. Another traditional way [73] might use Generative Adversarial Networks (GANs) based image to image translation methods in an unpaired way, such as CycleGAN [151] and UNIT [84], to transfer daytime images to fake nighttime images. Paired daytime and nighttime images are hard to obtain in the real-world applications, due to the dynamic traffic and environment changes. This kind of image translation considers this problem as domain adaptation for model fine-tuning on synthetic nighttime images without labeling the nighttime data. However, these methods also need extra time for model fine-tuning. In addition, GAN based image translation suffers from model collapse and does not preserve content details very well

[93, 84, 151, 54, 121]. Bottleneck layers in a general deep generator hurt the learning ability of convolution kernels due to downsampling and upsampling operations, resulting in possible losing some structure details. Besides, unpaired training data of different domains limits the detail-preserving ability of generators due to the lack of pixel-wise correspondence.

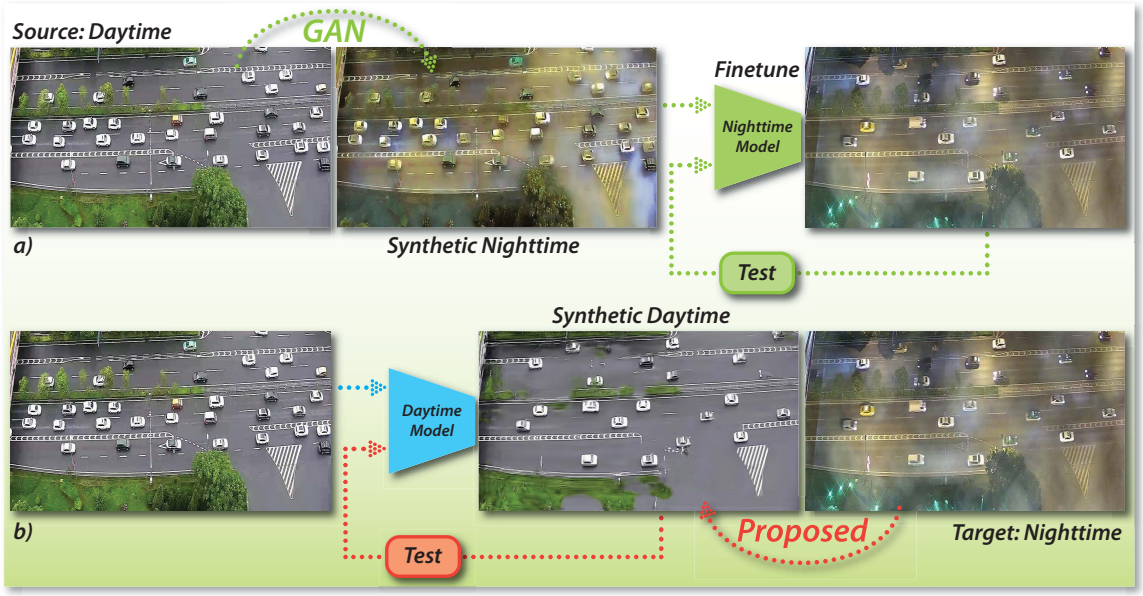

Figure 5.1: Illustration of the domain reuse problem: a) traditional method with style transfer and the nighttime model fine-tuned from the daytime model, b) the proposed detail-preserving Night-to-Day translation method without changing the daytime model.

In this chapter, we would like to reuse the daytime perception model to nighttime scenarios. Our basic idea is to maximally use the pre-trained daytime perception model, similar to works [133, 46], which could be easily extended to the nighttime tasks. Reversely to the traditional methods, we transfer the nighttime images back to the daytime style with the detail-preserving to reuse the trained daytime perception model, as shown in Fig. 5.3. The strengths of this reverse way are obvious and promising: 1) there are no extra training efforts for the already trained daytime perception model and no needs to manually label the nighttime data; 2) image transfer could reduce the domain distribution discrepancy between daytime and nighttime

data; 3) detail-preserving image transfer could better maintain the structure details than the GAN based image transfer.

Specifically, we propose a detail-preserving unpaired domain transfer method for this task, which mainly contains two components: 1) Style-transfer based StyleMix, 2) Kernel Prediction Network (KPN) based nighttime to daytime image transfer. Without paired daytime-nighttime image pairs, we propose to utilize style translation based StyleMix method, inspired by AugMix [49], to acquire pairs of daytime and nighttime images as training data for the following nighttime to daytime image transfer. We can effectively alleviate the detail corruption caused by GAN: 1) The synthetic nighttime image and corresponding daytime image translation can provide pixel-wise correspondence for night-to-day translation. 2) Kernel prediction network based method can refine the nighttime to daytime image translation because the per-pixel kernel fusion can effectively utilize the neighboring region for each pixel and could learn more spatial context representing structure information. The proposed method can conduct daytime and nighttime vehicle detection with just one daytime model, which is more convenient in real-world applications.

In this chapter, we choose the vehicle detection problem in the traffic surveillance video as a case study for the proposed approach. The KPN network is trained with object detection task together to adapt the trained daytime model to fit nighttime domain directly. Experimental results on a vehicle video dataset in daytime and nighttime verified the accuracy and effectiveness of the proposed approach. The contributions of this chapter are summarized in the following:

- We propose a detail-preserving unpaired domain transfer method for nighttime vehicle detection to adapt the trained daytime model directly for nighttime vehicle detection.
- To solve the problem of lack of paired daytime-nighttime image pairs, we propose to utilize style translation based StyleMix method to obtain pairs of daytime image

and nighttime image as training data. These training data are utilized by KPN network to perform nighttime to daytime image transfer.

- The comprehensive experimental results on a vehicle detection dataset from the video surveillance scenario in daytime and nighttime show that the proposed method achieved better vehicle detection performance in nighttime scenario.

5.2 Methodology

In this section, we propose the *detail-preserving unpaired domain transfer* for performing high accuracy object detection in the nighttime without retraining the detectors on daytime dataset. We first introduce the whole framework and reveal the challenges. Then, our two main contributions, *i.e.*, *scene-aware pixel-wise filtering* and *StyleMix*, help to address the challenges and achieve much better detection accuracy.

Detail-preserving Unpaired Domain Transfer for Nighttime Object Detection

We propose to perform nighttime object detection by transferring input nighttime images to the corresponding daytime versions for further object detection. This task could be simply formulated as

$$\hat{\mathbf{I}} = \phi(\mathbf{I}), \tag{5.1}$$

where the $\phi(\cdot)$ denotes a transfer function that can map the nighttime image $\mathbf{I}$ to the corresponding daytime version. A straightforward way is to set $\phi(\cdot)$ as a popular generator that can be trained with the adversarial loss. Nevertheless, we argue that GAN-based transfer is hard to recover the details in the nighttime image, which is rather important for accurate object detection. As shown in Fig. 5.2, the GAN-based method might destroy the detailed car structure, leading to missing detection.

Actually, the night-to-day translation for object detection requires that the object-related details, *e.g.*, car's structure, should be preserved while different scene patterns in the night should be perceived and properly mapped to its daytime versions, posing two challenges for deep learning-based solutions: ❶ Popular deep generator based methods easily harm the object details due to the common existing bottleneck layers where the input image is transferred by downsampling and upsampling. ❷ It is hard to get paired dataset which is significantly important for training detail-persevering networks with pixel-wise correspondence.

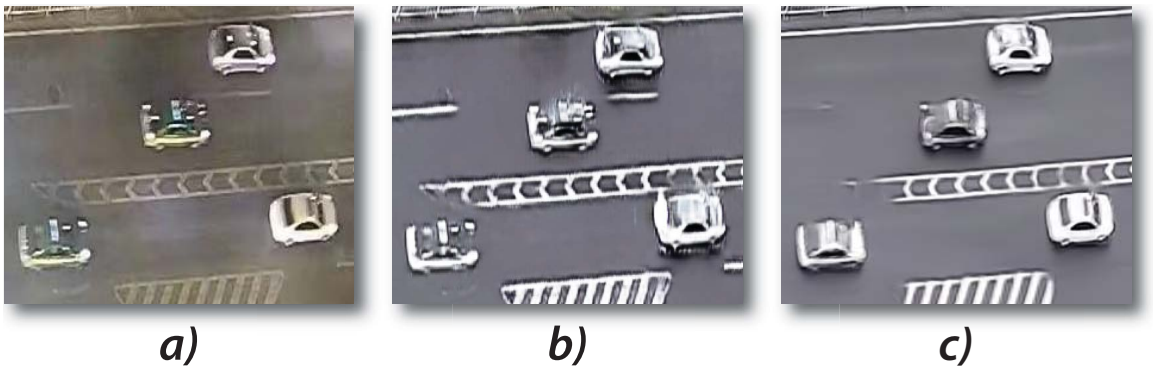

Figure 5.2: Image translation results of GAN-based method and the proposed method: a) target nighttime image, b) translated daytime image of a) by GAN-based method *CycleGAN* [151], c) translated daytime image of a) by the proposed method.

To address the first issue, we propose the *scene-aware pixel-wise filtering* for night-to-day transformation. Different to all existing works that employ a DNN as the transformer directly, our method maps the input image through a single-layer filtering whose kernels are predicted by an offline-trained DNN denoted as the *kernel prediction network*. Note that, the single-layer filtering (without any downsampling and upsampling operations) avoids the risk of missing important object-related details. Meanwhile, the DNN helps understand the scene and predict spatial-variant kernels for effective transformation. Specifically, kernel prediction network predicts a kernel for each pixel to capture local spatial context information to preserve more

details, *e.g.*, structure information. Recent works [6, 90, 40, 26] have proved that per-pixel kernel prediction network can achieve image recovery with better details. To address the second challenge, we propose a style-transfer-based data augmentation method, *i.e.*, *StyleMix*, to generate nighttime-daytime image pairs for training the kernel prediction network.

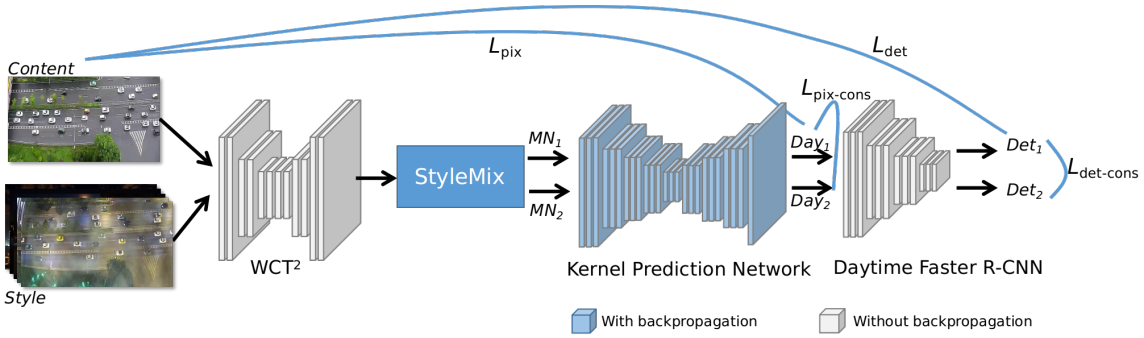

Figure 5.3: The proposed object detection pipeline at night with Night-to-Day image translation.

We show the whole framework in Fig. 5.3. Intuitively, our method is a pre-process module transferring the input nighttime image to daytime version for further object detection, which is supported by a novel and simple data augmentation method, *i.e.*, StyleMix.

Scene-aware Pixel-wise Filtering

We propose the scene-aware pixel-wise filtering for the night-to-day transformation. Specifically, we reformulate Eq. (5.1) as

$$\hat{\mathbf{I}} = \mathbf{K} \circledast \mathbf{I}, \tag{5.2}$$

$$\text{with} \quad \mathbf{K} = \phi(\mathbf{I}), \tag{5.3}$$

where $\circledast$ denotes the pixel-wise filtering, $\mathbf{K}$ is a pixel-wise filter $\in R^{(k\times k)\times h\times w}$. Each vector in channel dimension $\mathbf{K}(i,j) \in R^{(k\times k)}$ is a per-pixel kernel and can be applied to the $k \times k$ neighborhood region of each pixel in the input nighttime image $\mathbf{I}$ by

element-wise multiplication. The $\phi(\cdot)$ denotes the kernel prediction network and is used to perceive the input image and predict the suitable kernel for each pixel.

Then, we acquire the daytime version $\hat{\mathbf{I}}$ of the input image. Since it is pixel-wise filtering of the input nighttime image, it could largely preserve the image details without corruption. To fully leverage rich neighborhood information of every image pixel, a large kernel size k is desired, however, the computational and memory cost will increase as well. The kernel size k in our implementation is set to 5.

The framework of kernel prediction network is shown in Fig. 5.4. The training input data for KPN is synthetic nighttime images obtained from StyleMix. Specifically, two synthetic Mixed Nighttime images MN_1 and MN_2 with different style conditions are fed into KPN, respectively. KPN will output image-specific per-pixel filter for each image, respectively. Then element-wise multiplying the specific filter with the corresponding input image will generate the daytime version image $\hat{\mathbf{I}}_i$, $i = 1, 2$.

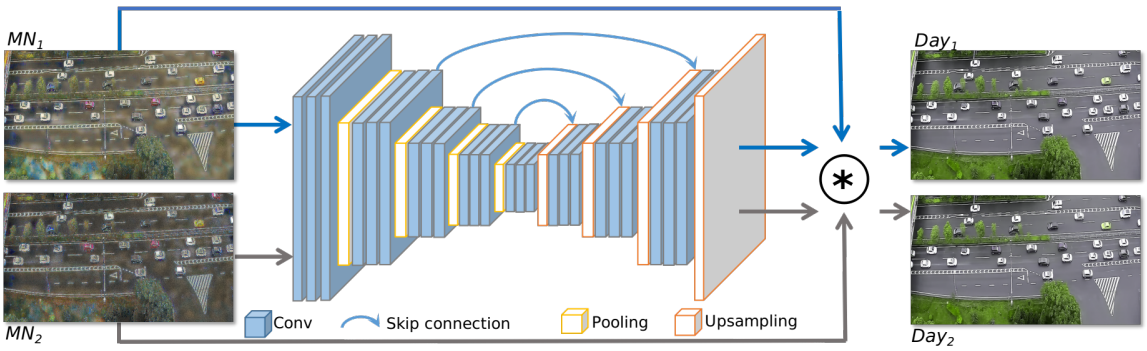

Figure 5.4: Illustration of the kernel prediction network based scene-aware pixel-wise filtering.

The basic loss function $\mathcal{L}_{\text{pix}}(\hat{\mathbf{I}}_i, \hat{\mathbf{I}}^*)$ is the pixel-wise L_1 distance between the ground truth daytime image $\hat{\mathbf{I}}^*$ and the translated daytime image $\hat{\mathbf{I}}_i$. It is defined as

$$\mathcal{L}_{\text{pix}}(\hat{\mathbf{I}}_i, \hat{\mathbf{I}}^*) = \|\hat{\mathbf{I}}^* - \hat{\mathbf{I}}_i\|_1. \tag{5.4}$$

We also define a consistency loss $\mathcal{L}_{\text{pix-cons}}$ between $\hat{\mathbf{I}}_1$ and $\hat{\mathbf{I}}_2$ by measuring their L_1

distance. The equation is

$$\mathcal{L}_{\text{pix}-\text{cons}}(\hat{\mathbf{I}}_1, \hat{\mathbf{I}}_2) = \|\hat{\mathbf{I}}_1 - \hat{\mathbf{I}}_2\|_1. \tag{5.5}$$

StyleMix: Bridging the Gap to Nighttime Data

The style-transfer-based method is utilized to generate nighttime-daytime image pairs for KPN training. To bridge the shift of synthetic nighttime and target nighttime data, we propose the SytleMix strategy to embody the diversity of nighttime scenarios.

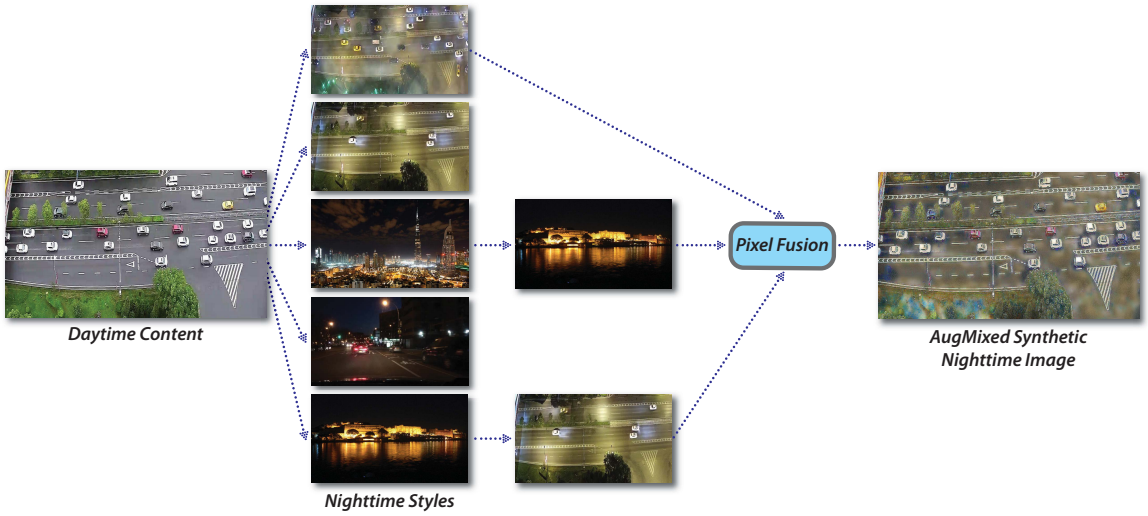

Figure 5.5: Illustration of the proposed StyleMix method to bridge the gap to nighttime data.

Specifically, style-transfer network can preserve the structure of input content image and stylize the content image according to the input style reference to implement image translation. As shown in Fig. 5.3, we adopt a pretrained style transfer network, whitening and coloring transforms (WCT^2), to finish daytime to nighttime image translation. For the input of WCT^2, daytime images are the content image and five real nighttime images act as the style reference. Five style reference images for following StyleMix are selected depending on the illumination condition of target nighttime scenarios. During daytime to nighttime image translation, StyleMix

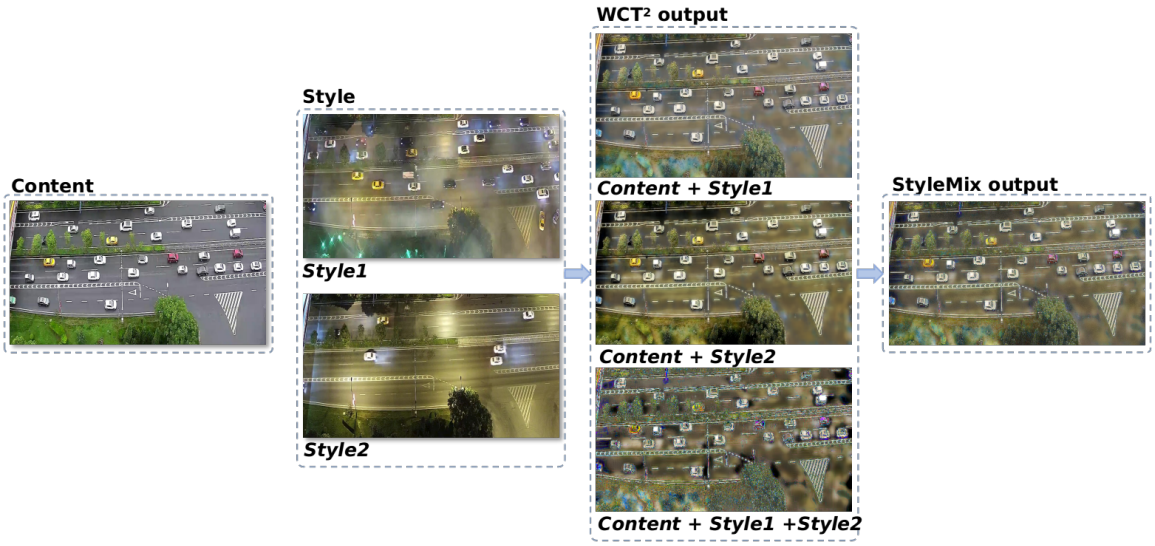

Figure 5.6: Sample visualization of the proposed StyleMix to generate synthetic nighttime images from real daytime images.

is involved to reduce the distribution shift of translated style and target nighttime style. It works in the way shown in Fig. 5.5. Specifically, for each daytime input image, three style augmentation chains out of fiver style reference are randomly sampled, each of which consists of one to two randomly selected style transfer operations. Then, the transferred images from these style augmentation chains are combined by pixel-level fusion to acquire a mixed nighttime image MN_i. Pixel-wise fusion is implemented by pixel-wise convex operations between translated nighttime images and convex coefficients. We randomly sample from a Dirichlet $(\alpha, \ldots, \alpha)$ distribution to construct the 3-dimensional vector of pixel-wise convex coefficients. Figure 5.6 shows one example of the pixel-wise fused synthetic nighttime image by StyleMix. The aug-mixed output of StyleMix are the pixel-wise fusion result of translated content image with different styles. It shows that StyleMix can effectively generate various kinds of synthetic nighttime images which are visually close to the real nighttime scene.

For detection task of the pipeline, we construct the detection loss $\mathcal{L}_{\mathrm{det}}(Det_i, Det^*)$ and the detection consistency loss $\mathcal{L}_{\mathrm{det-cons}}(Det_1, Det_2)$. We adopt Smooth L_1 loss [33] to calculate $\mathcal{L}_{\mathrm{det}}$ and $\mathcal{L}_{\mathrm{det-cons}}$. The total loss $\mathcal{L}_{\mathrm{N2D}}$ of the pipeline is a weighted

sum of $\mathcal{L}_{\text{pix}}$, $\mathcal{L}_{\text{pix-cons}}$, $\mathcal{L}_{\text{det}}$, and $\mathcal{L}_{\text{det-cons}}$. It is defined as

$$\mathcal{L}_{\text{N2D}} = \mathcal{L}_{\text{pix}} + \mathcal{L}_{\text{pix-cons}} + \lambda(\mathcal{L}_{\text{det}} + \mathcal{L}_{\text{det-cons}}), \tag{5.6}$$

where λ is set to 10 in our experiments.

5.3 Experiment

Datasets

In this chapter, the public D&N-Car Benchmark [73] is utilized to verify the effectiveness of the proposed approach. It is a real traffic surveillance dataset in urban expressway scene recorded in the city Xi'an, China. This dataset includes 1,200 daytime images and 1,000 nighttime images with their ground truth in the format of bounding boxes across different periods and dates, each of which is with resolution 1,280×720. There are total 57,059 vehicle instances in this dataset. The training set consists of 1,000 daytime traffic images with manual ground-truth labels, denoted as **Day-training**. The testing set includes 1,200 images, where 200 images are in daytime and 1,000 images are in nighttime. In the 200 daytime testing images, 100 images are in the normal traffic condition, denoted as **Day-normal**, and the other 100 images are in the congested traffic condition, denoted as **Day-congested**. The left 1,000 images of testing set constitute 4 subsets of nighttime traffic images (denoted as **Night1**, **Night2**, **Night3**, **Night4**). The details of the benchmark are shown in Table 5.1. In the experiment, we denote the labeled daytime traffic images (Day-training) as the Source Domain **S**, and the unlabeled nighttime traffic images as the Target Domain **T**.

Experimental Setting

We conduct experiments on two different scenarios: 1). Detect the vehicles during daytime by Faster R-CNN [98] model trained on **Day-training**; 2). Detect the vehi-

Table 5.1: Details of the D&N-Car Benchmark [73].

	No. of images	No. of car instances	Time
Day-training	1000	32,456	19:10
Day-normal	100	3,173	19:00
Day-congested	100	4,539	14:30
Night1	250	7,322	21:30
Night2	250	5,554	21:30
Night3	250	1,738	23:50
Night4	250	2,277	00:20

cles during nighttime by trained Faster R-CNN model on **Day-training** after the proposed night-to-day image translation. The detailed experimental setting is as follows: 1) Scenario 1: We directly train a Faster R-CNN model on the dataset **Day-training** in a supervised way and test the images on **Day-normal** and **Day-congested**, respectively; 2) Scenario 2: For style-transfer-based StyleMix image translation aiming at acquiring pairs of daytime and nighttime images, we utilize 1,000 images in **Day-training** set and 5 style reference images for nighttime images synthesis and augmixing styles. For KPN-based night-to-day training, there are 2,000 augmixed nighttime images for training in each epoch. Next, predicted daytime images are fed into detection task to further fit the translated daytime image for object detection. For inference, the trained KPN operates image translation for real nighttime images (**Night1**, **Night2**, **Night3**, and **Night4**), and then trained daytime detection model tests on translated nighttime images for performance evaluation.

We set the method that directly tests on nighttime images with trained daytime model Faster R-CNN [98] as a baseline. We also compare the proposed method with unpaired image translation methods UNIT [84], CycleGAN [151], and GcGAN [24] combining with Faster R-CNN in both day-to-night and night-to-day directions. To train the image translation models, the training dataset for daytime is the **Day-training** set and the training set for nighttime is a combination of **Night1**, **Night2**,

Night3, and **Night4**.

We built our translation and detection pipeline in PyTorch. For object detection, we use ResNet50 as our backbone. For detection training, we utilize Stochastic Gradient Descent (SGD) to optimize our network and set the initial learning rate to 0.0001 and decay it after every 10 epochs. The experiments are conducted on a single NVIDIA GTX 1080Ti GPU. For night-to-day image translation training, we train KPN with SGD by setting the learning rate to 0.002 for 200 epochs on two Tesla V100 GPUs. For a comprehensive performance evaluation, the widely-used object detection metric mAP (mean average precision) is used for evaluating the vehicle detection results. For all the experiments, the performance evaluation uses a uniform threshold of 0.5 for the Intersection Over Union (IoU) between the predicted bounding boxes and ground truth.

Results on Benchmark

We first report the detection results of one-stage detector SSD [86] and two-stage detector Faster R-CNN [98] for Scenario 1, shown in Table 5.2. We can see that both of the mAP drop from about 99% to 88% when the traffic is congested. The congested situation increases the object detection difficulty, resulting in a lower detection performance compared to the uncrowded situation. Because there is not a clear difference in terms of mAP between SSD and Faster R-CNN, we choose Faster R-CNN as our baseline detector for following experiments.

Table 5.2: Daytime vehicle detection results.

mAP(%)	SSD	Faster R-CNN
Day-normal	99.05	99.01
Day-congested	88.35	88.57

Results Compared to Day-to-night Translation Methods

We compare the detection results of nighttime vehicle to other image translation methods in a day-to-night direction. According to Scenario 2, the proposed method performs vehicle detection on translated daytime images obtained from KPN by the daytime model. However, for comparison methods performing nighttime vehicle detection in a day-to-night direction, they require additionally training a nighttime model for nighttime vehicle detection, besides the daytime model. For example, taking CycleGAN as the day-to-night image translation method, we translate daytime images to fake/synthetic nighttime images in an unpaired way, and followed by training a Faster R-CNN$_n$ detector on such fake/synthetic nighttime images with the same annotations of daytime images. Then we test the trained model on the nighttime images for vehicle detection. The comparison results are shown in Table 5.3.

Table 5.3: Nighttime vehicle detection results based on day-to-night translation. Note that the Faster R-CNN$_n$ model is trained on the fake/synthetic nighttime images.

Method \ mAP(%)	**Night1**	**Night2**	**Night3**	**Night4**	**Mean**
Mean-BGS [75]	54.03	49.09	52.16	55.56	52.71
SSD [86]	74.06	73.78	84.02	87.00	79.71
Faster R-CNN [98]	74.84	74.05	85.63	87.05	80.39
Faster R-CNN$_n$ [98]+ UNIT$_{d2n}$ [84]	70.56	77.13	82.87	88.19	79.68
Faster R-CNN$_n$ [98]+ CycleGAN$_{d2n}$ [151]	79.39	80.72	88.72	89.66	84.62
Faster R-CNN$_n$ [98]+ GcGAN$_{d2n}$ [24]	**80.89**	84.20	83.92	87.55	84.14
Proposed	80.25	**84.81**	**93.20**	**92.94**	**87.80**

We compare the detection results in the form of mAP for each subset of nighttime traffic images and the mean mAP for all of them. Day-to-night image translation methods UNIT [84], CycleGAN [151], and GcGAN [24] perform better than or comparable to the baseline Faster R-CNN which directly tests on nighttime images with daytime model. Taking the dataset **Night4** as an example, the proposed method, based on night-to-day image translation without retraining one more model, achieves the highest 92.94% mAP, about 5.4% higher than Faster R-CNN$_n$ + GcGAN$_{d2n}$, 3.3% higher than Faster R-CNN$_n$ + CycleGAN$_{d2n}$, 4.8% higher than Faster R-CNN$_n$

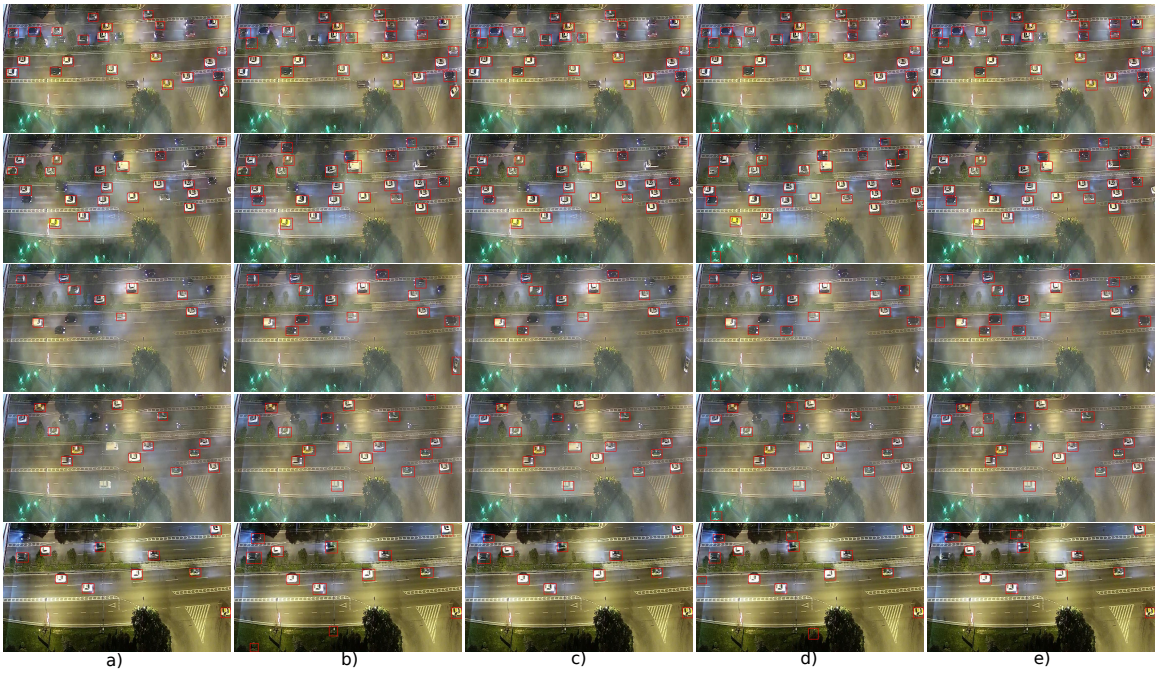

Figure 5.7: Visualization results of nighttime vehicle detection. a)-e) are the detection results from Faster R-CNN [98], Faster R-CNN$_n$ + UNIT$_{d2n}$ [84], Faster R-CNN$_n$ + CycleGAN$_{d2n}$ [151], Faster R-CNN$_n$ + GcGAN$_{d2n}$ [24], and the proposed method, respectively. Note: red bounding box indicates detection result.

+ UNIT$_{d2n}$ and 5.9% higher than the baseline Faster R-CNN. The proposed method achieves the best mean mAP with 87.80% for all nighttime traffic images, despite that Faster R-CNN$_n$ + GcGAN$_{d2n}$ performs a little better on the **Night1** subset. We also provide a traditional method Mean-BGS [75] performing vehicle detection through background subtraction and the daytime model of SSD [86] performing vehicle detection directly on nighttime images. Both of them are worse than Faster R-CNN for nighttime vehicle detection. As the corresponding detection results are shown in Fig. 5.7, we can clearly see that the proposed method is robust to various light conditions. UNIT, CycleGAN and GcGAN based methods could not well detect vehicles under poor light conditions and missed many black vehicles compared to the proposed method, and Faster R-CNN without any image translation does not perform well due to the domain shift of daytime and nighttime scenarios.

Results Compared to Night-to-day Translation Methods

We also compare our method with these image translation methods in a night-to-day direction. Comparison translation methods, UNIT$_{n2d}$ [84], CycleGAN$_{n2d}$ [151], and GcGAN$_{n2d}$ [24], first translate nighttime images to daytime-style images, and then these daytime-style images are fed into the daytime model for vehicle detection. The results are shown in Table 5.4. It shows that the proposed method could achieve the best mean mAP performance than UNIT, CycleGAN, and GcGAN for nighttime vehicle detection via night-to-day translation, demonstrating the advances of the proposed method. This is because the proposed method considers the per-pixel kernel fusion of neighboring information for each pixel and the object detection task during image translation training, preserving more features, *e.g.*, structure details, which are critical for the detection task.

Table 5.4: Nighttime vehicle detection based on night-to-day translation.

Method \ mAP(%)	**Night1**	**Night2**	**Night3**	**Night4**	**Mean**
Mean-BGS [75]	54.03	49.09	52.16	55.56	52.71
SSD [86]	74.06	73.78	84.02	87.00	79.71
Faster R-CNN [98]	74.84	74.05	85.63	87.05	80.39
Faster R-CNN [98]+ UNIT$_{n2d}$ [84]	62.54	62.86	81.29	81.00	71.92
Faster R-CNN [98]+ CycleGAN$_{n2d}$ [151]	77.69	79.12	88.31	88.79	83.47
Faster R-CNN [98]+ GcGAN$_{n2d}$ [24]	**83.57**	83.80	79.26	83.94	82.64
Proposed	80.25	**84.81**	**93.20**	**92.94**	**87.80**

Visualization results by the above mentioned image translation methods from nighttime to daytime are shown in Fig. 5.8. It shows that the proposed method could recover the daytime scenario with details. UNIT method suffers from model collapse presenting poor local textures and details. Specifically, the translated images are blurred, especially for the shape and edge of vehicles inside the image. It is consistent with the lower detection performance in the form of mAP in Table 5.4. CycleGAN could translate the texture of the vehicles from nighttime to daytime, but it is not robust to the intense road mirror reflections. It presents fake vehicles which do not

Figure 5.8: Visualization results of image translation from nighttime to daytime. a) is target nighttime image, b) to e) are the image translation results of UNIT$_{n2d}$ [84], CycleGAN$_{n2d}$ [151], GcGAN$_{n2d}$ [24], and proposed method, respectively.

exist in source nighttime images, resulting in more false positive detection samples. More black vehicles disappeared in the translated images by CycleGAN. GcGAN is also sensitive to such intense road mirror reflections, resulting in more fake vehicles in the translated images. Although the translated trees from the proposed method suffer from corruption, the goal of our work is to accurately detect vehicles at nighttime, we do not care much more about the tree corruption during the night-to-day translation. It is obvious that the translated cars are more natural with clear structures from the proposed method, resulting in better detection performance in Table 5.4. This is because the image translation training from nighttime to daytime makes full use of paired synthetic data with pixel-wise correspondence and per-pixel kernel fusion of neighboring information which provides rich spatial context information.

Table 5.5: Ablation study of the proposed method for the nighttime vehicle detection.

Method \ mAP(%)	**Night1**	**Night2**	**Night3**	**Night4**
Baseline	62.51	60.97	81.29	78.66
Baseline + StyleMix	75.90	76.20	88.20	86.90
Baseline + StyleMix + Zero	79.50	77.40	89.70	89.80
Baseline + StyleMix + Contrast	**80.20**	**81.79**	**92.11**	**91.58**

Ablation Study

In this section, we evaluate the contribution of each step in the proposed method: 1) training KPN without the StyleMix, instead, given each daytime image, we randomly select two style images from five style reference images to generate two synthetic nighttime images based on image translation, respectively. Two synthetic nighttime images are fed into KPN for training combined with detection task. We view this method as our *Baseline* in this section. 2) On the basis of step 1, we augment the style references in depth and width, denoted as *Baseline + StyleMix*. There are five style reference images for StyleMix, out of which two are from the nighttime traffic images of D&N-Car dataset. 3) For testing phase, we preprocess the target nighttime images by Zero-Reference network [36] to improve local contrast, then go through KPN for image translation to daytime scenario and followed by detection task via daytime detection model, denoted as *Baseline + StyleMix + Zero*. 4) Different with the preprocessing of step 4, we enhance the local contrast by improving the pixel value less than a threshold, denoted as *Baseline + StyleMix + Contrast*. The corresponding result of the proposed method in each step on the four night subsets in terms of mAP evaluation metric is shown in Table 5.5.

We can clearly see the positive effect of each step with the increasing mAP performance. Taking dataset **Night1** as an example, the baseline method could achieve 62.51% in mAP. When augmenting and mixing the style reference images to embody the diversity of synthetic nighttime scenarios, mAP increases by about 13%. When

prepossessing the target nighttime images with Zero-Reference network and Contrast, the mAP continues to increase to 79.50% and 80.20%, respectively.

We conduct ablation experiments to verify the effectiveness of the style reference setting for StyleMix. We construct a night-style image pool for style reference selection, which consists of 21 images, 7 from nighttime dataset of D&N-Car, 7 from BDD dataset [134] with nighttime scene and the other 7 from WCT^2 publicized project website[1]. We randomly choose 5 images from this night-style image pool as style reference for the whole image translation model. We conduct three experiments with different style reference setting for StyleMix. There are 1, 2, and 5 images from the nighttime images of D&N-Car dataset as different experiment settings: $StyleMix_1$, $StyleMix_2$ and $StyleMix_5$. The detection results are shown in Table 5.6. It turns out that the detection performance increases with more night-style images involved from the D&N-Car dataset. It is reasonable since we expect the StyleMix model to render the synthetic nighttime image closer to the corresponding nighttime style of nighttime images.

Table 5.6: Ablation study with different style reference setting.

Method \ mAP(%)	**Night1**	**Night2**	**Night3**	**Night4**	**Mean**
Baseline + $StyleMix_1$+ Contrast	76.27	79.51	91.76	91.69	84.80
Baseline + $StyleMix_2$ + Contrast	80.20	81.79	92.11	91.58	86.42
Baseline + $StyleMix_5$ + Contrast	**80.25**	**84.81**	**93.20**	**92.94**	**87.80**

5.4 Chapter Summary

In this chapter, we proposed a detail-preserving method to implement the nighttime to daytime image translation and thus adapt daytime trained detection model to nighttime detection. We firstly utilize style translation method to acquire paired images of daytime and nighttime, which are hard to obtain in real-world applications.

[1]https://github.com/clovaai/WCT2

We propose to stylemix the reference styles to embody the diversity of synthetic nighttime scenarios. The following nighttime to daytime translation is implemented based on kernel prediction network to avoid texture corruption and trained with detection task to make the translated daytime image not only visually photo-realistic to the daytime scenario but also fit the detection task to reuse the daytime domain knowledge. The proposed method can perform both daytime and nighttime vehicle detection with one model. Experimental results showed that the proposed method achieved effective and accurate nighttime detection results.

Chapter 6

Shadow removal for image visual quality enhancement and facial landmark detection improvement simultaneously

6.1 Overview

Facial landmark detection [128, 149, 61] is a fundamental step for numerous facial related applications, *e.g.*, face recognition and verification [153, 87], 3D face reconstruction [83], and safety-critical applications, *e.g.*, deepfake detection [150, 79], and facial reenactment [139, 110] for virtual avatar applications. While recent deep-learning techniques bring us continuously improved landmark-detection performance, most of them are designed to handle images of "clean faces". However, in real-world applications, face images usually contain image degradations, such as noise [9, 58], shadow [144, 51], and haze [91], which degrades the aesthetic quality of images directly and may further affect the performance of landmark detectors.

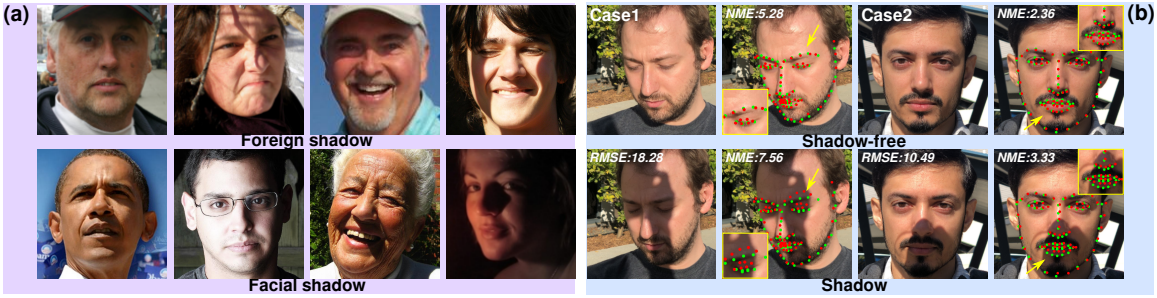

Figure 6.1: Illustrations of (a) various shadow scenes on facial landmark detection benchmarks [101, 126] and (b) effects of foreign shadow on image quality and facial landmark detection [22]. Red: prediction. Green: ground truth. RMSE measures the image degradation caused by shadow, and NME evaluates the detection error.

As a natural phenomenon, shadow is very common on face images – in practice, the light to any face region can be occluded by surrounding objects or by part of itself. This is especially true for portrait images captured in the wild with unconstrained environments. As shown in Fig. 6.1(a), portrait shadow happens in two kinds of scenes: foreign shadow and facial shadow. Foreign shadow appears when there is an external occluder (*e.g.*, a tree or a hat brim) blocking the light source to reach out to the subjects' face. The foreign shadow can present an arbitrary 2D shape in the natural image, relying on the shape of the occluder and position of the primary light source. In contrast, facial shadow casts on the face by the face itself due to the

facial geometry and presents a small space of 2D shapes when natural lighting is not perfectly uniform [139].

The foreign shadow effects on facial landmark tasks are under-explored, although there are works [144, 51] exploring illumination invariance for face recognition which cannot be simply extended to facial landmark detection. Such works are mainly designed for facial shadows caused by the intensity and position of the light source in the indoor scene. Note that foreign shadows are almost always distracting compared to facial shadows. Image intensity edges appear in most of foreign shadow scenes, which are uncorrelated to facial geometry and will obfuscate facial 3D structure. By contrast, the intensity edges introduced by facial shadows are more likely to be helpful for inferring the shape of face. Therefore, we aim to remove the foreign shadow entirely. In this chapter, we set the research scope to tackle foreign shadows. As a result of shadow cast, spatial-variant illumination and color distortion in the shadow region [26] degrade the image quality and undermine the image features significantly. As shown in Fig. 6.1(b), shadowed faces hurt the image quality with large root mean square error (RMSE), and presents unreasonable and much deteriorated landmark locations for the eyebrows (See Case1) and mouth (See Case2), as measured by much degraded NME scores.

An intuition way to alleviate the performance loss caused by shadow is to restore the underlying shadow-free image utilizing current state-of-the-art (SOTA) shadow removal methods. However, there are two challenges posing to such a solution: ❶ The interplay between light, occluder, and the subject directly affects the shadow appearance. As a result, in the real world, shadow patterns are significantly diverse, which increases the difficulty of shadow removal algorithms. ❷ Even though shadow removal methods could obtain high visual-quality images with lower RMSE, the landmark detection performance may even get worse compared to that of shadow images due to the potential domain shift between landmark detection and image

quality enhancement. Existing works (haze [94] and rain [50] removal) demonstrate that visual quality improvement benefits little or even hurts the high-level perception task performance. All above facts motivate us to answer two basic questions: how shadow affects the landmark detection, and whether shadow removal can benefit the robustness of landmark detectors.

To this end, for the first time, we propose to link the two seemingly independent but intrinsically related tasks, *i.e.*, shadow removal and facial landmark detection, by constructing a totally novel dataset and benchmark. Such a solution has never been tried in both communities before this work. Note that, constructing such a benchmark is challenging and not trivial since the shadow patterns are not exhaustive, and *existing facial landmark detection benchmarking techniques* [101, 126] *collecting natural images cannot meet the requirements*: ❶ There are about less than 2% of data in each benchmark [101, 126] presenting foreign shadow scenes. ❷ For those foreign shadow samples, as shown in Fig. 6.1(a), though with increasing shadow intensity, they primarily exhibit less abrupt edges and their shadow patterns are limited.

To alleviate these challenges, we propose novel solutions to ensure the comprehensiveness: ❶ We employ the physical model of shadow and synthesize facial shadow images by considering four common factors (*i.e.*, intensity, size, shape, and location) with three severities, ❷ We investigate the shadow from the perspective of adversarial attack and propose a totally new attack (*i.e.*, adversarial shadow attack) to identify shadow patterns that are more challenging to landmark detection. ❸ We introduce a real-world shadow face dataset for verifying the generalization ability of facial landmark detectors. With these elaborated designs, we are able to quantitatively and systematically study the effect of shadows on the facial landmark detection.

Moreover, we study whether shadow removal can help improve the robustness of landmark detectors covering three SOTA shadow removal methods and three SOTA landmark detectors. We observe that shadow removal can not only improve the

image visual quality, but also boost the performance of landmark detection – there is a positive correlation between the shadow-removal accuracy and the landmark detection accuracy. Note that, such a relationship is not apparent in haze-removal and classification task [94] or even becomes opposite in deraining and detection task [50]. In this chapter, the relationship is dominant especially when shadow degradation level is higher (*i.e.*, higher-severity shadow and adversarial shadow). It implies that feature embedding spaces of shadow removal and landmark detection aiming to optimize partially overlap with each other, which provides a bridge for the two tasks. Inspired by this observation, we further propose a new shadow-removal framework regularized by landmark detection to further improve the visual quality and landmark detection simultaneously.

Overall, we summarize our contributions as follows:

- We construct a new shadow-face benchmark SHAREL, including synthetic shadow-face dataset, adversarial shadow-face dataset, and real shadow-face dataset, by comprehensively considering shadow intensity, size, shape, and locations and developing a novel adversarial attack.
- Based on SHAREL, we comprehensively and quantitatively study the effects of shadow and shadow removal on image visual quality and the performance of facial landmark detection.
- We propose a novel shadow removal framework with awareness of facial landmark detection and verify its performance on the proposed benchmark, boosting both the shadow removal and landmark detection performance.

Overview

Natural shadow presents diverse shadow patterns in the wild due to the influences of occluders and light sources. For example, different light occluders can lead to diverse shadow appearances with different sizes and shapes. In addition, the illumination level, material of occluders and object surface where shadow casts determine the reflection and scattering of the light, which may affect the intensity at the shadow region. Nevertheless, enumeration of all permutations formulating patterns is not practical due to dynamic and complex scenes. To alleviate this issue and analyze the effects of shadow and shadow removal on facial landmark detection extensively, we propose three dataset construction strategies: ❶ We follow the well-known and widely used physical shadow model to synthesize shadowed faces on the clean facial landmark detection dataset (*i.e.*, 300W [101]) and consider four factors (*i.e.*, intensity, size, shape, and location) with three severities. ❷ To mine hard shadow images that affect landmark detection easily, we think this problem from the perspective of adversarial attack and propose a novel synthesis method (*i.e.*, *adversarial shadow attack*). ❸ To address the potential shifting problem between synthesized shadow faces and the real ones, we also introduce 100 real shadow face images as a subset of the whole dataset. We present examples for the three strategies in Fig. 6.2 and detail each strategy in the following.

Synthetic Shadowed Faces

Physical model of shadow. We adopt the well-known and widely used physical model of shadow in [106]. Specifically, following the illumination and reflectance formulation of an image [106], we can represent a clean (*i.e.*, shadow-free) image

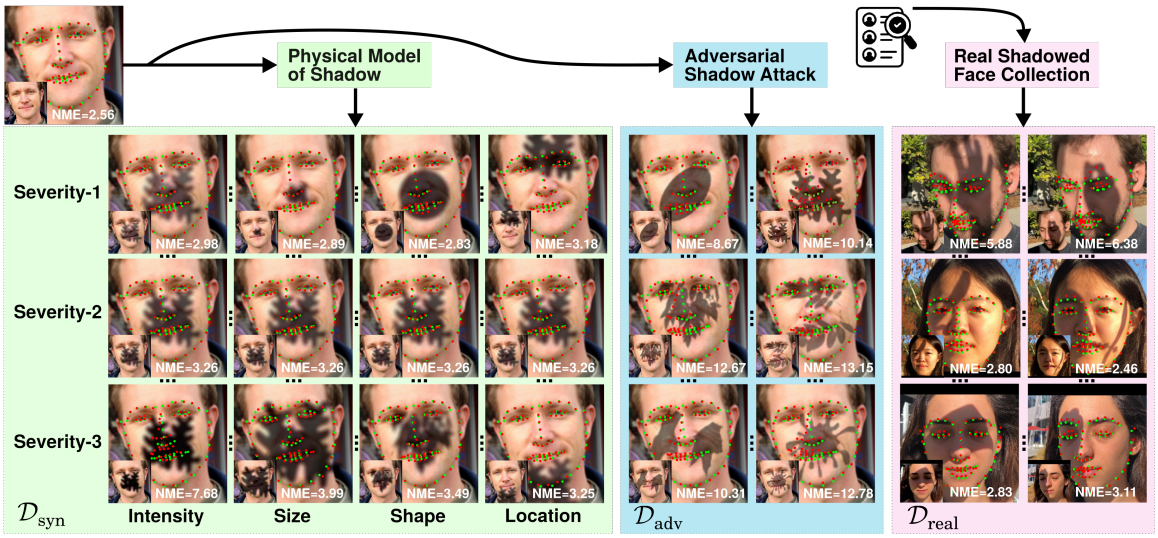

Figure 6.2: Three dataset construction strategies including physical model-based synthesis, adversarial shadow attack, and real shadowed face collection. Green: ground truth. Red: prediction. NME measures the landmark detection performance. The lower, the better.

captured under a single primary light source as

$$\mathbf{I}_p^{\text{cln}} = \mathbf{L}_p \mathbf{R}_p = (\mathbf{L}_p^{\text{d}} + \mathbf{L}_p^{\text{a}})\mathbf{R}_p, \tag{6.1}$$

where $\mathbf{I}_p^{\text{cln}}$, $\mathbf{L}_p$, and $\mathbf{R}_p$ are pixel intensity, illumination, and reflectance at the p-th pixel, respectively. The illumination stems from two sources, *i.e.*, the direct illumination $\mathbf{L}^{\text{d}}$ and the ambient illumination $\mathbf{L}^{\text{a}}$. When an occluder appears in front of the light source, the direct illumination disappears while the ambient illumination is also affected. We can represent the p-th shadowed pixel as

$$\mathbf{I}_p^{\text{shd}} = \alpha \mathbf{L}^{\text{a}} \mathbf{R}_p = \alpha(\mathbf{I}_p^{\text{cln}} - \mathbf{L}_p^{\text{d}} \mathbf{R}_p), \tag{6.2}$$

where α is a scalar and determines the attenuation of the ambient illumination, which is caused by the occluder. With a clean image $\mathbf{I}^{\text{cln}}$ and a dark image $\mathbf{I}^{\text{shd}}$, we can represent an image $\mathbf{I}$ containing a shadow region, following [145, 56], as

$$\mathbf{I} = \mathbf{I}^{\text{shd}} \odot \mathbf{M} + \mathbf{I}^{\text{cln}} \odot (1 - \mathbf{M}), \tag{6.3}$$

where $\mathbf{M}$ is a binary map that defines the shadow region and is determined by the occluder. To generate more realistic shadow, we reformulate Eq. (6.3) to

$$\mathbf{I} = \mathbf{I}^{\text{shd}} \odot \rho(\mathbf{D} \odot \mathbf{M}) + \mathbf{I}^{\text{cln}} \odot (1 - \rho(\mathbf{D} \odot \mathbf{M})), \tag{6.4}$$

where ρ models light scattering and spatial variation and $\mathbf{D}$ is a face depth map. Note that, the images of living faces have face-like depth information, which are critical for anti-spoofing application. $\mathbf{D}$ can make generated shadow more realistic, which is not considered in previous shadow models [145, 56]. Moreover, to generate realistic shadow pattern, we borrow the implementation in [47, 145] and use the function ρ to render the depth-aware mask (*i.e.*$\mathbf{D} \odot \mathbf{M}$) to become a shadow matte image by modeling the light scattering beneath human skin and modeling the spatial variation of the shadow via a spatially-varying blur. Please find more details in [145].

Then, we can substitute Eq. (6.2) into Eq. (6.4) and get

$$\mathbf{I} = \text{ Shadow}(\mathbf{I}^{\text{cln}}, \mathbf{M}, \alpha) = (1 - (1 - \alpha)\rho(\mathbf{D} \odot \mathbf{M}))\mathbf{I}^{\text{cln}} + \alpha\beta\rho(\mathbf{D} \odot \mathbf{M}), \tag{6.5}$$

where $\beta = -\mathbf{L}^{\text{d}}\mathbf{R}$ representing the response of the camera to the reflected direct illumination and the ambient attenuation α does not depend on the light source (*e.g.*, wavelength) [106]. Moreover, as demonstrated and discussed in [56], β is a three-channel vector and can be estimated from the α via a linear transformation.

Overall, *given a clean face image* $\mathbf{I}^{\text{cln}}$*, a shadow map* $\mathbf{M}$*, a depth map* $\mathbf{D}$*, and the* α*, we can synthesize a shadowed face* $\mathbf{I}$. In practice, we use the 3DDFA-V2 [37] to predict the depth map from the clean image.

Synthesized shadows with different factors and severities. To cover extensive shadow patterns in the real world, we generate shadowed faces for a clean face image from four factors: intensity, size, shape, and location.

i. *Intensity.* The illumination level and material of object surfaces determine the reflection and scattering of light, resulting in shadow with diverse intensities. We

model the shadow intensity via the parameter α in Eq. (6.5) since it directly models the relationship between shadowed pixels and illuminated pixels. α is about in range [0.0, 1.0) for realistic shadow scene [56]. We uniformly sample α from ranges $[0.8, 1.0)$, $[0.4, 0.6)$, $[0.0, 0.2)$, for light, medium and heavy shadows. The lower α, the heavier the shadow. For different shadow intensity level design, we want to quantify how much texture and content degradation shadow brings, and how that affects visual quality and landmark detection. We present three kinds of intensities for the same face in Fig. 6.2.

ii. *Size.* The size of an occluder blocking the light and position of the light source directly affect the area of the shadow (*i.e.*, shadow size). We model shadow size via the number of non-zero pixels in $\mathbf{M}$ in Eq. (6.5) and consider three different severities, *i.e.*, small, medium, and large shadow regions. Intuitively, large-size shadow will degrade image quality more than small-size shadow because face-related information (*e.g.*, structure) becomes less. Given a specified shadow shape, we can set the shadow areas (*i.e.*, number of non-zero pixels in $\mathbf{M}$) to take up $10\% \sim 20\%$, $45\% \sim 55\%$, and $80\% \sim 90\%$ areas of the face images by rescaling the shadow region in $\mathbf{M}$, which corresponds to three severities, *i.e.*, small, medium, and large shadow regions. We show the three different shadow sizes for the same face in Fig. 6.2.

iii. *Shape.* Occluders with different 3D geometrical shapes and the lights with different positions relative to the same occluder also affect the shadow shapes. We represent the shadow shape via the shadow mask in $\mathbf{M}$ in Eq. (6.5). To cover diverse shadow shapes, we collect a silhouette dataset containing 132 shapes of natural objects, and classify them into three levels by a shape complexity metric defined in [11], which is denoted as E. The shape complexity metric considers two aspects during measurement, *i.e.*, the distance distribution of the contour points of a shape to its centroid and the smoothness of the contour. Intuitively, if the complexity of a shape is low, the shape may tend to be a circle or has smooth contour. We present three

shapes for the same face in the Fig. 6.2, their complexity values are 0.04, 0.10, and 0.15 from severity 1 to 3. With the collected silhouette dataset, we first calculate the shape complexity for each collected shape. Then, we sort all shapes according to the complexity and evenly divide them into three severities, *i.e.*, low, medium, and high complexities.

iv. *Location.* We further consider the shadow position in the face image, since the shadow with the same intensity, size, and shape may still have different effects to shadow removal and landmark detection methods. For example, facial landmarks include clues of eyebrows, eyes, nose, jaw, and mouth. Shadow degradation to different parts of the facial structure will help quantitatively recognize the importance of each structural information to landmark detection. We model the shadow location via the centroid position of the shadow mask in $\mathbf{M}$ and consider three scenarios with the position at top, middle, and bottom of the whole face. For implementation, we split the whole face into three parts, *i.e.*, top, middle, and bottom regions. Then, we shift the centroid point of shadow mask in $\mathbf{M}$ to the center of the three regions. Specifically, if the center point of the whole face is $(\frac{W}{2}, \frac{H}{2})$ where H and W are the height and width of the face, the center points of the top, middle, and bottom regions are $(\frac{W}{2}, \frac{H}{6})$, $(\frac{W}{2}, \frac{H}{2})$, and $(\frac{W}{2}, \frac{5H}{6})$, respectively.

Synthetic shadowed face subset $\mathcal{D}_{\mathbf{syn}}$. With the above synthesis strategies, given a clean face image, we can generate three shadowed faces for each factor, which corresponds to three severities. To consider the effects of all factors, we have $3^4 = 81$ shadowed faces across all factors and severities for each clean image. Then, based on the facial landmark dataset 300W [101] that contains 689 clean face images for testing landmark detectors, we can generate a larger dataset with $81 \times 689 = 55,809$ shadowed images. We present some examples in Fig. 6.2. Although the constructed dataset covers diverse shadow patterns, it cannot represent all possible situations, in

particular, the hard cases that SOTA landmark detectors cannot address. To alleviate this issue, we propose a novel adversarial attack based on the physical model of shadow to mine the hard shadow patterns.

Adversarially Shadowed Faces

Given an image, adversarial attack is to calculate an imperceptible noise-like perturbation under the guidance of a targeted deep model, and then add it to the image. As a result, the corrupted image can mislead the targeted model easily. Unlike traditional adversarial attacks based on additive perturbations, recently there is a growing trend in developing non-additive adversarial attacks that enjoy better transferability and stealthiness such as blur-based adversarial attacks [39, 38, 55], attacks based on weather elements [136, 28] and lighting conditions [30, 112, 111, 12], and other modalities [29, 122, 41, 77], *etc.* We can regard the adversarial attack as a way to mine hard noise patterns that cannot be addressed by the targeted deep model. Here, we propose a novel attack method, *i.e.*, *adversarial shadow attack*, and further extend it to generate hard shadow patterns that are able to fool the landmark detectors. Therefore, we can evaluate the shadow robustness.

By intuition, we can tune the physical parameters of shadow model, as described in Shadow($\mathbf{I}^{\text{cln}}, \mathbf{M}, \alpha$) in Eq. (6.5), like the α and $\mathbf{M}$ under the supervision of landmark detectors to cover different shadow patterns with different intensities, sizes, shapes, and locations. Specifically, given a clean face image $\mathbf{I}^{\text{cln}}$ and a pre-trained landmark detector $\varphi(\cdot)$ we want to evaluate, we can: 1) First use Eq. (6.5) to synthesize the shadowed image, and feed it to $\varphi(\cdot)$. 2) We get the detection results and calculate the loss according to the ground truth (*i.e.*, $\mathbf{y}$). 3) We tune the physical variables $\mathbf{M}$ and α iteratively to maximize the landmark detection loss. As a result, the synthesized face can fool the detection easily while maintaining the physical properties of the

shadow. We can formulate the above process by

$$\begin{aligned} &\underset{\mathbf{M},\alpha,\vartheta}{\operatorname{argmax}} \mathcal{J}(\varphi(\text{Shadow}(\mathbf{I}^{\text{cln}}, \text{Aff}_{\vartheta}(\mathbf{M}), \alpha)), \mathbf{y}), \\ &\text{subject to } \ \|\mathbf{M} - \mathbf{M}_0\|_{\text{p}} < \epsilon_{\text{M}}, \|\alpha - \alpha_0\| < \epsilon_{\alpha}, \|\vartheta - \vartheta_0\|_{\text{p}} < \epsilon_{\vartheta}, \end{aligned} \tag{6.6}$$

where $\mathcal{J}(\cdot)$ is the loss function of landmark detection. Note that, different from the raw synthesis function in Eq. (6.5), $\text{Shadow}(\mathbf{I}^{\text{cln}}, \text{Aff}_{\vartheta}(\mathbf{M}), \alpha)$ conducts the affine transformation (*i.e.*, $\text{Aff}_{\vartheta}(\cdot)$) on $\mathbf{M}$ before feeding it for synthesis, which allows us to mine more shadow shapes with a given shadow mask. The ϑ contains six affine parameters. Like general adversarial attack methods, we set the L_{p} norm to $\mathbf{M}$, α, and ϑ to force the optimization space within a ball of ϵ_{M}, ϵ_{α}, and ϵ_{ϑ}, around their initialization (*i.e.*, $\mathbf{M}_0$, α_0, and ϑ_0), respectively.

To solve the Eq. (6.6), we follow the general adversarial attack methods: ❶ We set $\mathbf{M}_0$, α_0, and ϑ_0, and get the initial synthesized image. ❷ We feed the generated image to the landmark detector $\varphi(\cdot)$ and calculate the loss. ❸ We conduct back-propagation and get the gradients of $\mathbf{M}$, α, and ϑ w.r.t.the loss function. ❹ We calculate the sign of the gradients and use them to update the three variables by multiplying the gradients with three step sizes. ❺ We generate a new synthesized image and loop step-2 to step-4 for a number of iterations. In terms of the initialization, we select $\mathbf{M}_0$ from the collected 132 silhouette images and set α_0 to be 0.8. Then, We initialize ϑ_0 as $\begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \end{bmatrix}$, $\text{Aff}_{\vartheta}(\mathbf{M}) = \mathbf{M}$ during initialization. In terms of the implementation, we set the step size of α, ϑ, and $\mathbf{M}$ as 0.01, 0.02, and 0.0012, respectively. The number of iterations is set to be 40. we use ∞ norm for L_{p}, and set ϵ_{α}, ϵ_{ϑ}, and ϵ_{M} as 0.4, 0.8, and 0.0.048, respectively. As a result, adversarial shadow images can present more hard shadow patterns against landmark detectors, as shown in Fig. 6.2, the NMEs in $\mathcal{D}_{\text{adv}}$ could be over 10 compared to around 3 in $\mathcal{D}_{\text{syn}}$.

Adversarially shadowed face subset $\mathcal{D}_{\mathbf{adv}}$. With the above method, given a landmark detector and the 300W dataset, we first conduct attack for each image, and then evaluate the detector on the adversarially shadowed faces. Thus, for each detector, we have an exclusive new version of 689 adversarially shadowed face images to evaluate their robustness.

Real Shadowed Faces

Real shadowed face subset $\mathcal{D}_{\mathbf{real}}$. To verify the shadow effect on visual quality and landmark detection performance in the real-world scenario, we introduce a real-world shadow portrait dataset [145]. However, this dataset lacks facial landmark annotations for landmark detection evaluation. We first obtain pseudo ground truth by a SOTA pre-trained HRNet [118], and then refine it manually as the final landmark ground truth. Finally, we have 9 subjects and 100 pairs of shadowed and shadow-free portrait images captured in the outdoor scenes with varied face poses, shadow shapes, and illumination conditions. Figure 6.2 presents some examples.

6.3 Shadow Removal & Landmark Detection Benchmark (SHAREL)

Setups

Datasets. Our main data is constructed based on the landmark detection benchmark 300W [101]. 300W contains $3,148$ clean face images for training and 689 clean images for testing. Each image is labeled with 68 landmarks. We construct SHAREL-based on the testing dataset of 300W. We add shadow patterns to the 300W and get $\mathcal{D}_{\text{syn}}$; we propose adversarial shadow attack and obtain $\mathcal{D}_{\text{adv}}$ for each landmark detector; we collect real shadowed faces (*i.e.*, $\mathcal{D}_{\text{real}}$) to further enrich our dataset. Finally, our dataset $\{\mathcal{D}_{\text{syn}}; \mathcal{D}_{\text{adv}}; \mathcal{D}_{\text{real}}\}$ has $\{55,809; 689; 100\}$ shadowed and shadow-free image pairs (total $56,598$ pairs) that are labeled with 68 landmarks.

We additionally construct $\{\mathcal{D}^{\mathrm{t}}_{\mathrm{syn}}, \mathcal{D}^{\mathrm{t}}_{\mathrm{adv}}\}$ from a randomly selected subset (1,500 clean images) of 300W training set for training shadow removal models. Each of $\{\mathcal{D}^{\mathrm{t}}_{\mathrm{syn}}, \mathcal{D}^{\mathrm{t}}_{\mathrm{adv}}\}$ contains 1,500 shadow-free and shadowed image pairs. For creating $\mathcal{D}^{\mathrm{t}}_{\mathrm{syn}}$, each clean image uniformly selects a severity for each factor to generate the shadow image. $\mathcal{D}^{\mathrm{t}}_{\mathrm{adv}}$ follows the same shadow generation way of $\mathcal{D}_{\mathrm{adv}}$.

Metrics. To clarify the shadow and deshadow effect on image quality, we adopt the Root Mean Square Error (RMSE) metric in LAB color space for evaluation, similar to [26, 69, 53]. For facial landmark detection evaluation, we adopt Normalized Mean Error (NME) metric with inter-ocular distance as normalization strategy following [66, 22, 118]. Both the lower, the better.

Evaluated methods. With our SHAREL, we can evaluate the quality restoration capability of the shadow removal methods and the detection accuracy of facial landmark detectors on different shadow or deshadowed patterns. We first analyze three SOTA facial landmark detectors, *i.e.*, SAN [22], HRNet [118], and LUVLi [66], under different shadow patterns. All landmark detectors are pre-trained on clean face images. Further, we utilize three SOTA deep shadow removal methods, *i.e.*, MaskShadow-GAN [53], SP+M-Net [69], and AEFNet [26], to handle the shadowed faces in SHARELand discuss whether and how these methods can help improve landmark detection performance. All shadow removal algorithms are trained on dataset $\mathcal{D}^{\mathrm{t}}_{\mathrm{syn}}$ and $\mathcal{D}^{\mathrm{t}}_{\mathrm{adv}}$ separately for fair comparison, and shadow removal models trained on $\mathcal{D}^{\mathrm{t}}_{\mathrm{syn}}$ are also utilized to test on real data.

Evaluation Results and Discussion

Effects of shadow to image quality and facial landmark detection. In Fig. 6.3(a-c), we report the RMSEs of shadow images and landmark detection results with NMEs in $\{\mathcal{D}_{\mathrm{syn}}, \mathcal{D}_{\mathrm{adv}}, \mathcal{D}_{\mathrm{real}}\}$ to identify the shadow degradation on image quality and

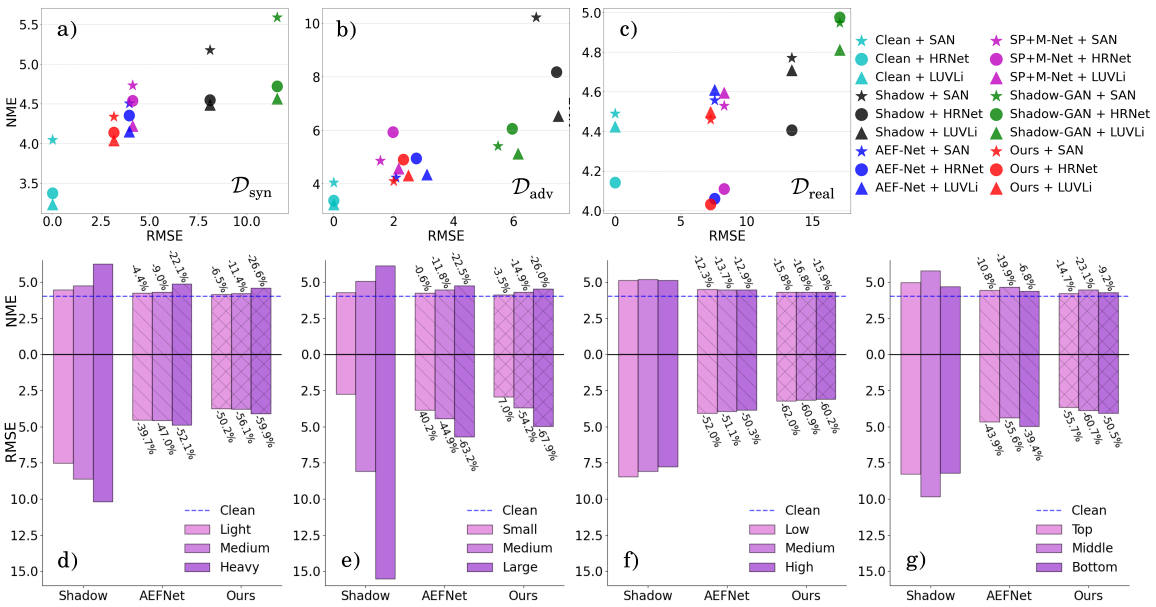

Figure 6.3: Shadow removal and landmark detection performance on SHAREL. (a-c): shadow removal (RMSE) and landmark detection (NME) results of $\{\mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{adv}}, \mathcal{D}_{\text{real}}\}$ subsets, respectively. Each color represents results on shadow-free images (*e.g.*, Clean+*), shadow images (*i.e.*, Shadow+*), and shadow-removed images with four shadow removal methods (*e.g.*, AEFNet/SP+M-Net/MaskShadow-GAN/Ours+*). Different icon shapes represent different landmark detectors. (d-g): shadow pattern analysis of landmark detection (NME) and shadow removal (RMSE) results of $\mathcal{D}_{\text{syn}}$ for **intensity** (d), **size** (e), **shape** (f), and **location** (g). Blue dash line represents the result on clean images by the pre-trained landmark detector SAN [22]. Each group represents results on shadow images (*i.e.*, Shadow), and shadow-removed images with two shadow removal methods (*e.g.*, AEFNet/Ours). Each color represents a severity type. Relative performance gain, *i.e.*, the percent of NME/RMSE drops, after shadow removal compared to shadow images is listed for AEFNet and Ours.

detection performance. Figure 6.3(d-g) report the shadow pattern analysis on $\mathcal{D}_{\text{syn}}$ with four factors. The detector adopted in(d-g) is SAN [22]. The results show that: ❶ Compared with shadow-free images, shadow images have high RMSEs since the shadow harms the image quality significantly. More intense the shadow degradation, worse the visual quality. For example, the RMSE of shadow and shadow-free images of large-size with 15.52 is higher than that of small-size with 2.74 in $\mathcal{D}_{\text{syn}}$ (Fig. 6.3e). Intensity, size, and location, instead of shape, are dominant factors affecting the shadow degradation. ❷ According to the NME results, we observe that: the

performance of all landmark detectors drops when shadow appears in images and hard shadow pattern, *i.e.*, higher-severity shadow and adversarial shadow, hurts the detection task most. Specifically, the landmark detector SAN [22] achieves 4.05 NME on clean images of $\mathcal{D}_{\text{adv}}$, while the NME of shadow images increases by 152.3% to 10.22 (Fig. 6.3b). In $\mathcal{D}_{\text{syn}}$, heavy-intensity shadow achieves 6.26 NME with 54.7% performance drop compared to NME of clean images, while the performance loss caused by light-intensity shadow is 10.2% by SAN [22] (Fig. 6.3d).

In summary, shadow hurts the image quality and landmark detection significantly. Higher-severity presents high degradation capacity, that is, two tasks suffer from larger performance loss with increasing RMSEs and NMEs. The same performance loss trend appears in the image quality and landmark detection.

Effects of shadow removal to image quality enhancement and facial landmark detection. We perform shadow removal on shadow images, and present RMSEs and NMEs of shadow-removed images in $\{\mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{adv}}, \mathcal{D}_{\text{real}}\}$ to evaluate the effectiveness of shadow removal methods. The results are shown in the Fig. 6.3. We can observe that: ❶ Shadow removal methods present different capabilities on the image quality enhancement (Fig. 6.3(a-c)). To be specific, SP+M-Net [69] and AEFNet [26] can enhance the image quality significantly in all subsets. MaskShadow-GAN [53] further hurts the quality in the subsets $\{\mathcal{D}_{\text{syn}}, \mathcal{D}_{\text{real}}\}$ while achieving counterpart result in $\mathcal{D}_{\text{adv}}$. The former mainly stems from that MaskShadow-GAN, *i.e.*, a GAN-based image translation method, introduces artifacts during training. The reason why MaskShadow-GAN performs better on $\mathcal{D}_{\text{adv}}$ may be that shadow pattern generated by MaskShadow-GAN overlaps with that of $\mathcal{D}_{\text{adv}}$. Specifically, during MaskShadow-GAN training, it also generates diverse shadow patterns taking unpaired shadow-free images and shadow masks as input, and such shadow-pattern images are not covered by normal shadow images via a discriminator in an adversarial training way, simi-

lar to the adversarial generation process of $\mathcal{D}_{\text{adv}}$. ❷ Higher-severity shadow pattern achieves much larger relative gain for image quality enhancement. For example, large-size shadow-removed images acquire 63.2% visual quality improvement compared to 40.2% quality degradation of small-size shadow in $\mathcal{D}_{\text{syn}}$ (Fig. 6.3e). The latter further quality degradation stems from the over smoothing of current shadow removal methods. In addition, $\mathcal{D}_{\text{adv}}$ also achieves much larger gain with 69.3% compared to 51.2% of $\mathcal{D}_{\text{syn}}$ by SAN (Fig. 6.3(a-b)). ❸ The same performance gain trend of SOTA shadow removal methods and higher-severity shadow pattern presents in the landmark detection evaluation. In Fig. 6.3e, the large-size shadow pattern obtains the highest 22.5% NME decreasing compared to 0.6% of small-size shadow. The $\mathcal{D}_{\text{adv}}$ achieves 58.5% performance improvement compared to 13.0% of $\mathcal{D}_{\text{syn}}$ by SAN (Fig. 6.3(a-b)).

In summary: ❶ *Current SOTA shadow removal methods can effectively improve the image quality and landmark detection simultaneously.* ❷ *Higher-severity achieves much larger performance gain after shadow removal for image quality and landmark detection.* ❸ *There is a positive correlation between shadow removal and landmark detection tasks. To be specific, landmark detection performance decreases with degraded image quality caused by shadow and improves with increasing image quality after shadow removal. In particular, when image quality suffers from higher degradation,* i.e.*, higher-severity shadow and adversarial shadow, the performance gain trend keeps consistent for the two tasks after shadow removal.*

Note that, such positive correlation does not always exist in computer vision tasks. For example, deraining even hurts the object detection performance on rainy images [50]. Haze-removal improves classification task with very limited margin [94]. The positive correlation between shadow removal and landmark detection implies that the embedding spaces they optimized somehow overlap with each other. However, previous shadow removal works [53, 69, 26] only focus on recovering pleasing visual

images, ignoring the mutual influence between them. We propose a novel framework to explore the mutual influence of the two tasks to verify whether they can benefit from each other.

6.4 Landmark-regularized Shadow Removal

To link the facial landmark detection and shadow removal, we propose to introduce the landmark detection embedding to regularize the shadow removal by the *mutual attention fusion module*. Moreover, we propose extra regularization loss functions by jointly considering the image reconstruction and landmark detection, as shown in Fig. 6.4.

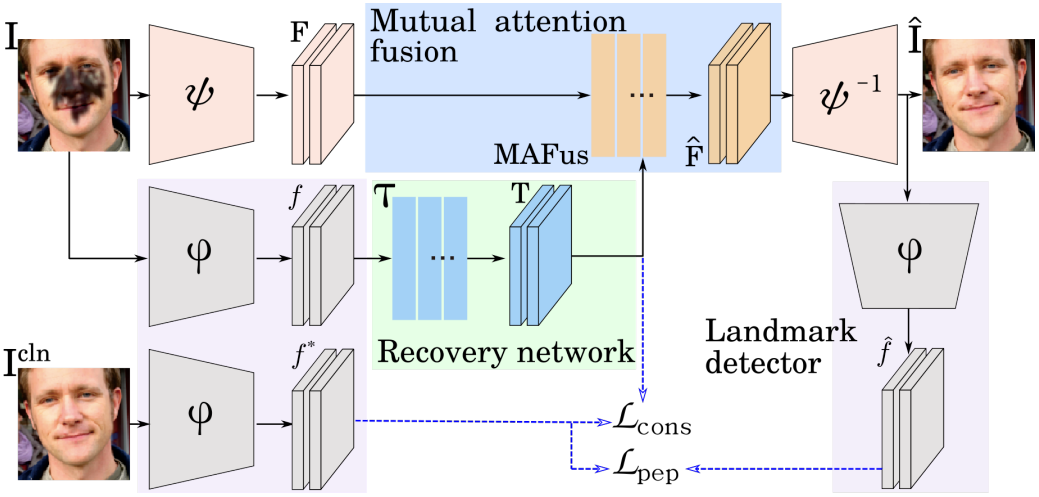

Figure 6.4: Landmark-regularized shadow removal network.

Given a shadow image $\mathbf{I}$, a shadow removal method can be generally represented as

$$\hat{\mathbf{I}} = \psi^{-1}(\mathbf{F}), \text{and } \mathbf{F} = \psi(\mathbf{I}), \tag{6.7}$$

where $\psi(\cdot)$ and $\psi^{-1}(\cdot)$ are the encoder and decoder of the shadow removal model, respectively, and $\hat{\mathbf{I}}$ is the shadow-removed image. Considering a backbone model $\varphi(\cdot)$ of a landmark detector, we propose to fuse the embeddings of the shadow removal

model and landmark detection model, and modify Eq. (6.7) as

$$\hat{\mathbf{I}} = \psi^{-1}(\hat{\mathbf{F}}), \text{and } \hat{\mathbf{F}} = \text{MAFus}(\psi(\mathbf{I}), \tau(\varphi(\mathbf{I}))), \tag{6.8}$$

where $\tau(\cdot)$ is a recovery network to align the embedding of $\varphi(\mathbf{I})$ to the space of $\varphi(\mathbf{I}^{\text{cln}})$ for better detection-aware embedding. MAFus$(\cdot)$ is the mutual attention fusion module and can leverage the features from different networks and integrate information complementary.

Mutual attention fusion (MAFus). Inspired by the recent self-mutual attention based on non-local module [124] for fusing multiple modalities [85], we employ it for embedding fusion. Specifically, we first map the input features $\mathbf{F} = \psi(\mathbf{I})$ and $\mathbf{T} = \tau(\varphi(\mathbf{I}))$ to three spaces like non-local network

$$\theta^{\text{f}}(\mathbf{F}) = \mathbf{F}\mathbf{W}_{\theta}^{\text{f}}, \phi^{\text{f}}(\mathbf{F}) = \mathbf{F}\mathbf{W}_{\phi}^{\text{f}}, g^{\text{f}}(\mathbf{F}) = \mathbf{F}\mathbf{W}_{g}^{\text{f}}, \tag{6.9}$$

$$\theta^{\text{t}}(\mathbf{T}) = \mathbf{T}\mathbf{W}_{\theta}^{\text{t}}, \phi^{\text{t}}(\mathbf{T}) = \mathbf{T}\mathbf{W}_{\phi}^{\text{t}}, g^{\text{t}}(\mathbf{T}) = \mathbf{T}\mathbf{W}_{g}^{\text{t}}. \tag{6.10}$$

Then, we can use the Gaussian function for self similarity calculation on θ^* and ϕ^* spaces

$$f^{\text{f}}(\mathbf{F}) = \theta^{\text{f}}(\mathbf{F})\phi^{\text{f}}(\mathbf{F})^{\top}, f^{\text{t}}(\mathbf{T}) = \theta^{\text{t}}(\mathbf{T})\phi^{\text{t}}(\mathbf{T})^{\top}. \tag{6.11}$$

After that, we calculate the mutual attention based on the two similarity results

$$\mathbf{A}^{\text{f}}(f^{\text{f}}(\mathbf{F}), f^{\text{t}}(\mathbf{T})) = \text{softmax}(f^{\text{f}}(\mathbf{F}) + \gamma^{\text{t}} f^{\text{t}}(\mathbf{T})), \tag{6.12}$$

$$\mathbf{A}^{\text{t}}(f^{\text{f}}(\mathbf{F}), f^{\text{t}}(\mathbf{T})) = \text{softmax}(f^{\text{t}}(\mathbf{F}) + \gamma^{\text{f}} f^{\text{f}}(\mathbf{T})), \tag{6.13}$$

where γ^{t} and γ^{f} are pixel-wise attention weights to fuse embedding attentions, and are predicted by the concatenation of $\mathbf{F}$ and $\mathbf{T}$. With the mutual attention result $\mathbf{A}^{\text{f}}$ and $\mathbf{A}^{\text{t}}$, we can obtain the non-local outputs of $\mathbf{F}$ and $\mathbf{T}$,

$$\mathbf{Z}^{\text{f}} = (\mathbf{A}^{\text{f}} g^{\text{f}}(\mathbf{F}))\mathbf{W}_{z}^{\text{f}} + \mathbf{F}, \tag{6.14}$$

$$\mathbf{Z}^{\text{t}} = (\mathbf{A}^{\text{t}} g^{\text{t}}(\mathbf{T}))\mathbf{W}_{z}^{\text{t}} + \mathbf{T}. \tag{6.15}$$

The final output of MAFus is the concatenation of $\mathbf{Z}^{\mathrm{f}}$ and $\mathbf{Z}^{\mathrm{t}}$, *i.e.*, $\hat{\mathbf{F}} = [\mathbf{Z}^{\mathrm{f}}, \mathbf{Z}^{\mathrm{t}}] = \mathrm{MAFus}(\mathbf{F}, \mathbf{T})$.

Loss functions. We employ L_1 distance for the image reconstruction loss $\mathcal{L}_{\mathrm{pix}}(\hat{\mathbf{I}}, \mathbf{I}^{\mathrm{cln}}) = \|\mathbf{I}^{\mathrm{cln}} - \hat{\mathbf{I}}\|_1$. We propose three more regularization loss functions to explore the detection embedding guidance for shadow removal, which are detection regularization loss $\mathcal{L}_{\mathrm{det}}$, detection-aware perceptual loss $\mathcal{L}_{\mathrm{pep}}$, and detection-aware consistency loss $\mathcal{L}_{\mathrm{cons}}$.

Detection regularization loss aims to provide a regularization item for constraining the shadow removal process to satisfy the landmark detection. The weights of pre-trained landmark detector with clean images are fixed and only shadow removal is optimized. Given a shadow-removed image $\hat{\mathrm{I}}$, detection embedding $\hat{f}$ and heatmap $\hat{h}$ can be inferred by $\hat{f}, \hat{h} = \varphi(\hat{\mathbf{I}})$. For its corresponding shadow-free image, $f^*, h^* = \varphi(\mathbf{I}^{\mathrm{cln}})$. The detection regularization loss is defined as $\mathcal{L}_{\mathrm{det}}(\hat{h}, h^*) = \mathrm{MSE}(\hat{h}, h^*)$. Moreover, inspired by perceptual loss [60], we align the detection embeddings of shadow-removed and clean images by $\mathcal{L}_{\mathrm{pep}}(f^*, \hat{f}) = \mathrm{MSE}(f^*, \hat{f})$. Finally, we propose detection-aware consistency loss $\mathcal{L}_{\mathrm{cons}}$ to align the embeddings of $\tau(\varphi(\mathbf{I}))$ and $\varphi(\mathbf{I}^{\mathrm{cln}})$. The $\mathcal{L}_{\mathrm{cons}}$ aims to drive transformed detection embedding of shadow image to that of shadow-free image, their consistency renders the $\tau(\varphi(\mathbf{I}))$ to provide rich complementary information for better shadow removal guidance. It is formulated to $\mathcal{L}_{\mathrm{cons}}(f^*, \mathbf{T}) = \mathrm{MSE}(f^*, \mathbf{T})$. The total loss of the proposed framework is

$$\mathcal{L} = \mathcal{L}_{\mathrm{pix}} + \lambda_1 \mathcal{L}_{\mathrm{det}} + \mathcal{L}_{\mathrm{cons}} + \lambda_2 \mathcal{L}_{\mathrm{pep}}, \tag{6.16}$$

where λ_1 and λ_2 are set to 0.1 and 10 in our experiments.

6.5 Experiment

Setups. We conduct extensive experiments to verify our proposed landmark-regularized shadow removal method. Based on baseline method AEFNet [26], we cumulatively

add each module for contribution evaluation: 1) Detection regularization loss $\mathcal{L}_{\text{det}}$. We adopt SAN [22] as the weight-fixed landmark detector. 2) Mutual attention fusion module MAFus. A shadow image is fed into SAN and the output detection feature map will be directly fused with shadow removal feature map via MAFus. 3) Detection-aware consistency loss $\mathcal{L}_{\text{cons}}$. Detection feature map of shadow image is fed into the recovery network τ for feature alignment, followed by performing MAFus. 4) Detection-aware perceptual loss $\mathcal{L}_{\text{pep}}$. Feature shift of shadow-removed image and clean image is further optimized by $\mathcal{L}_{\text{pep}}$. The training set and testing set are $\mathcal{D}^{\text{t}}_{\text{syn}}$ and $\mathcal{D}_{\text{syn}}$. Results are shown in Table 6.2.

Implementing details. The proposed pipeline is implemented in PyTorch. We build our proposed framework based on shadow-removal method AEFNet [26] and training setting keeps the same with official publicized code of AEFNet.

Recovery network (τ). Given a shadow image $\mathbf{I}$ and a landmark detector $\varphi(\cdot)$, we can obtain $f, h = \varphi(\mathbf{I})$, where $f \in R^{N\times C\times H\times W}$ is the detection embedding from the landmark detector and h is the output heatmap representing landmark localization. N, C, H, and W are batch size, channel number, height, and width of f, respectively. We extract f from the third convolution block of the backbone of landmark detector SAN [22] with respective field 32×32. Then, f is fed into the recovery network $\tau(\cdot)$ for embedding alignment. The architecture of the network τ is listed in Table 6.1. The outputs of convolution layers "Conv1_2", "Conv2_2", and "Conv3_2" will be concatenated and go through "Conv1_fuse" for the final output. Multi-level feature fusion is designed for less information loss. The output of the recovery network is $\mathbf{T} = \tau(\varphi(\mathbf{I}))$, where $\mathbf{T}, \in R^{N\times C\times H\times W}$, has the same dimension with f.

Mutual attention fusion weights of MAFus. Given the embedding of shadow removal model $\mathbf{F} = \psi(\mathbf{I})$ and aligned landmark detection embedding $\mathbf{T} = \tau(\varphi(\mathbf{I}))$, we can obtain two similarity results $f^{\text{f}}(\mathbf{F})$ and $f^{\text{t}}(\mathbf{T})$ based on non-local module [124].

Table 6.1: The architecture of recovery network.

	Input Channel	Output Channel	Filter Size	Stride	Pad
Conv1_1	256	256	3	1	1
Conv1_2	256	256	3	1	1
Conv2_1	256	128	3	1	1
Conv2_2	128	128	3	1	1
Conv3_1	128	64	3	1	1
Conv3_2	64	64	3	1	1
Conv_fuse	448	256	1	1	0

Previously, attention for each embedding is calculated based on its similarity result only. In contrast, mutual attention is achieved by a weighted sum of attentions of different modalities. The mutual attentions of shadow removal embedding $\mathbf{A}^{\mathrm{f}}$ and landmark detection embedding $\mathbf{A}^{\mathrm{t}}$ are calculated by

$$\mathbf{A}^{\mathrm{f}}(f^{\mathrm{f}}(\mathbf{F}), f^{\mathrm{t}}(\mathbf{T})) = \mathrm{softmax}(f^{\mathrm{f}}(\mathbf{F}) + \gamma^{\mathrm{t}} f^{\mathrm{t}}(\mathbf{T})), \tag{6.17}$$

$$\mathbf{A}^{\mathrm{t}}(f^{\mathrm{f}}(\mathbf{F}), f^{\mathrm{t}}(\mathbf{T})) = \mathrm{softmax}(f^{\mathrm{t}}(\mathbf{F}) + \gamma^{\mathrm{f}} f^{\mathrm{f}}(\mathbf{T})), \tag{6.18}$$

where γ^{t} and γ^{f}, $\in R^{N\times 1\times H\times W}$, are pixel-wise attention weights. Both H and W are 32. γ^{t} and γ^{f} are predicted by a convolution layer followed by a BN [57] layer and the ReLU activation function. The kernel size of convolution layer is 1. The input to the convolution layer is the concatenation of $\mathbf{F}$ and $\mathbf{T}$, *i.e.*, [$\mathbf{F}$, $\mathbf{T}$]. $\mathbf{F}$ has the same dimension with $\mathbf{T}$.

Table 6.2: Ablation study of shadow removal and landmark detection results on the $\mathcal{D}_{\mathrm{syn}}$ dataset. "Proposed" means $\{\mathcal{L}_{\mathrm{det}}$, MAFus, $\mathcal{L}_{\mathrm{cons}}$, $\mathcal{L}_{\mathrm{pep}}\}$.

Methods	Shadow removal / RMSE			Landmark detection / NME		
	Shadow	Non-shadow	All	SAN [22]	HRNet [118]	LUVLi [66]
Clean	0.00	0.00	0.00	4.04	3.37	3.24
w/ shadow	33.27	0.53	8.09	5.17	4.55	4.49
AEFNet [26]	9.14	2.39	3.95	4.50	4.35	4.15
+ $\mathcal{L}_{\mathrm{det}}$	7.56	1.93	3.23	4.40	4.18	4.08
+ MAFus	7.16	1.96	3.16	4.35	4.15	4.06
+ $\mathcal{L}_{\mathrm{cons}}$	7.18	1.93	3.14	4.34	4.14	4.05
+ $\mathcal{L}_{\mathrm{pep}}$	7.09	1.97	3.15	4.33	4.13	4.04
SP+M-Net [69]	11.26	1.97	4.11	4.73	4.54	4.22
+ Proposed	7.58	1.96	3.28	4.34	4.16	4.06

Results and discussion. In Table 6.2, taking SAN detector as an example, it turns out that: 1) $\mathcal{L}_{\text{det}}$ improves the shadow removal capacity with 18.2% decreasing RMSE compared to baseline in the whole image. Landmark detection also benefits from it with 2.2% decreasing NME. Similar results are obtained in training denoising network [82] via regularization by semantic segmentation task. High-level vision information can provide guidance for the image reconstruction process. 2) The attention-based feature fusion MAFus and $\mathcal{L}_{\text{cons}}$ further reduce the RMSE to 3.16 and 3.14 in the whole image. Correspondingly, landmark detector also performs better with reaching 4.35 NME and 4.34 NME, respectively. 3) With $\mathcal{L}_{\text{pep}}$, the proposed method performs best in the shadow region with 7.09 RMSE for image quality evaluation and with 4.33 NME for landmark detection evaluation. In summary, compared to shadow images, the proposed method improves the visual quality by 78.7% in the shadow region, and increases the landmark detection performance of SAN by 16.2%. Same detection performance improvement trend presents for all detectors. In Fig. 6.3(c), our proposed method even achieves better detection performance on shadow-removed images compared to on clean images by SAN.

Results under different shadow patterns. We conduct shadow pattern analysis on $\mathcal{D}_{\text{syn}}$ considering different factors. When evaluating the shadow and shadow removal effects by one factor to image quality and facial landmark detection, we enumerate other factors for a comprehensive analysis. For example, given the 689 clean face images, when evaluating the intensity factor with slight severity, we collect $27 \times 689 = 18,603$ shadowed images with slight intensity while diverse sizes, shapes, and locations. Given different shadow patterns, we analyze shadow and shadow removal effects to image quality and landmark detection performance covering four shadow removal methods, *i.e.*, MaskShadow-GAN [53], SP+M-Net [69], AEFNet [26], and Ours and three landmark detectors, *i.e.*, SAN [22], HRNet [118], and LUVLi [66].

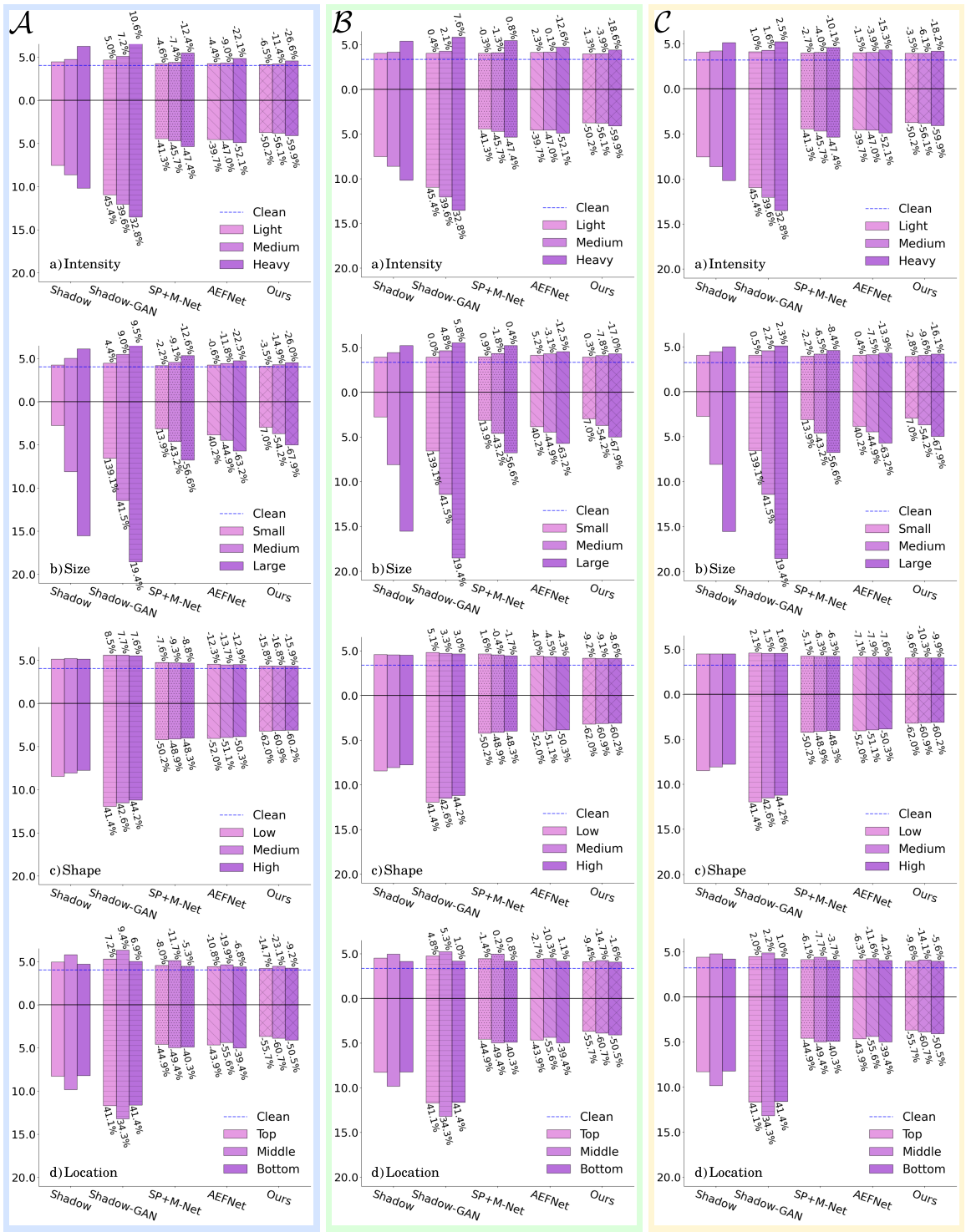

Figure 6.5: Shadow pattern analysis of shadow removal and landmark detection performance on $\mathcal{D}_{\text{syn}}$. ($\mathcal{A}$-$\mathcal{C}$): shadow removal (RMSE) and landmark detection (NME) results with shadow removal methods (*i.e.*, MaskShadow-GAN [53], SP+M-Net [69], AEFNet [26], and Ours) and detectors (*i.e.*, SAN [22] ($\mathcal{A}$), HRNet [118] ($\mathcal{B}$), and LUVLi [66] ($\mathcal{C}$)). (a-d): landmark detection and shadow removal results of $\mathcal{D}_{\text{syn}}$ for **intensity** (a), **size** (b), **shape** (c), and **location** (d).

The results are shown in Fig. 6.5. The blue dash line represents the result of clean images by the pre-trained landmark detectors. Each group along the x-axis represents results on shadow images (*i.e.*, Shadow), and shadow-removed images with four shadow removal methods (*e.g.*, MaskShadow-GAN/SP+M-Net/AEFNet/Ours). Each color represents a severity type. Relative performance gains, *i.e.*, the percent of NME/RMSE drops, after shadow removal compared to shadow images are listed for MaskShadow-GAN, SP+M-Net, AEFNet, and Ours. Note: Shadow-GAN denotes the MaskShadow-GAN.

It turns out that: 1) Shadow affects the image quality and facial landmark detection performance significantly with higher RMSE and NME. More intense the shadow degradation, worse the visual quality and landmark detection performance for all landmark detectors. For example, heavy-intensity shadow images achieve 5.13 NME by landmark detector LUVLi compared to 4.10 NME of slight-intensity shadow images (See Fig. 6.5-$\mathcal{C}$(a)). 2) Shadow removal can reduce the performance loss of the image quality and landmark detection caused by shadow. Higher severity achieves higher performance gain after shadow removal. For example, small-size shadow increases the performance loss by 0.3% after shadow removal by our method via landmark detector HRNet compared to 17.0% decreasing NME of large-size shadow (See Fig. 6.5-$\mathcal{B}$(b)). 3) For all detectors, shadow-removed images by MaskShadow-GAN obtain worse image quality and landmark detection performance due to the introduced artifacts. For example, medium-size shadow-removed images achieve 41.5% higher RMSE and 4.8% higher NME by HRNet (See Fig. 6.5-$\mathcal{B}$(b)). 4) Intensity, size, and location are dominant factors affecting shadow degradation, *i.e.*, various intensities, sizes, and locations have obviously diverse effects to image quality and facial landmark detection before and after shadow removal. For example, Figure 6.5(b) shows that shadow images with large size have higher RMSE (15.22) with shadow-free images compared to that of small size (2.74). However, the relative difference

of RMSE of shadow and shadow-free images between different shape severities is within 0.7. The trend is the same as the shadow effect on facial landmark detection performance with various detectors. After shadow removal, image quality and facial landmark detection performance are similar between various shape complexities with comparable RMSEs and NMEs.

Detection feature benefits shadow removal. Facial landmark detection features capture the main structure of a face, which is less sensitive to the intensity variation caused by shadow. By embedding the detection features in the shadow removal pipeline, we actually regularize the shadow removal features to let them preserve the main structure information and be less sensitive to the shadow either. To validate this, we calculate the pre-fused features (*i.e.*, $\mathbf{F}$ in Fig. 6.4) and fused features (*i.e.*, $\hat{\mathbf{F}}$) of a shadowed face and the ground truth face, respectively, and get $\mathbf{F}$, $\hat{\mathbf{F}}$, $\mathbf{F}_{\text{gt}}$, and $\hat{\mathbf{F}}_{\text{gt}}$. Then, we calculate the $L_1(\mathbf{F}, \mathbf{F}_{\text{gt}})$ and $L_1(\hat{\mathbf{F}}, \hat{\mathbf{F}}_{\text{gt}})$ for all examples in $\mathcal{D}_{\text{real}}$. The average $L_1(\mathbf{F}, \mathbf{F}_{\text{gt}})$ is 0.288 while the average $L_1(\hat{\mathbf{F}}, \hat{\mathbf{F}}_{\text{gt}})$ is 0.070. Clearly, after the landmark-aware fusion, the features become less sensitive to the shadow due to the much smaller L_1-distance.

Effectiveness of $\mathcal{L}_{\text{cons}}$ and $\mathcal{L}_{\text{pep}}$. Taking AEFNet+$\mathcal{L}_{\text{det}}$+MAFus for SAN as baseline in Table 6.3, we gradually add $\mathcal{L}_{\text{cons}}$ and $\mathcal{L}_{\text{pep}}$ to it. Both $\mathcal{L}_{\text{cons}}$ and $\mathcal{L}_{\text{pep}}$ contribute mainly for removing high-severity shadow. For example, shadow removal with large-size type improves a lot than that with small size for detection.

Table 6.3: Effects of $\mathcal{L}_{\text{cons}}$ and $\mathcal{L}_{\text{pep}}$.

Shadow size	baseline	+ $\mathcal{L}_{\text{cons}}$	+ $\mathcal{L}_{\text{pep}}$
Large	5.04	5.01	4.98
Small	2.90	2.89	2.90

Generalization to SP+M-Net. In Table 6.2, it shows that proposed landmark regularization pipeline can also help SP+M-Net [69] to boost both the shadow removal and landmark detection.

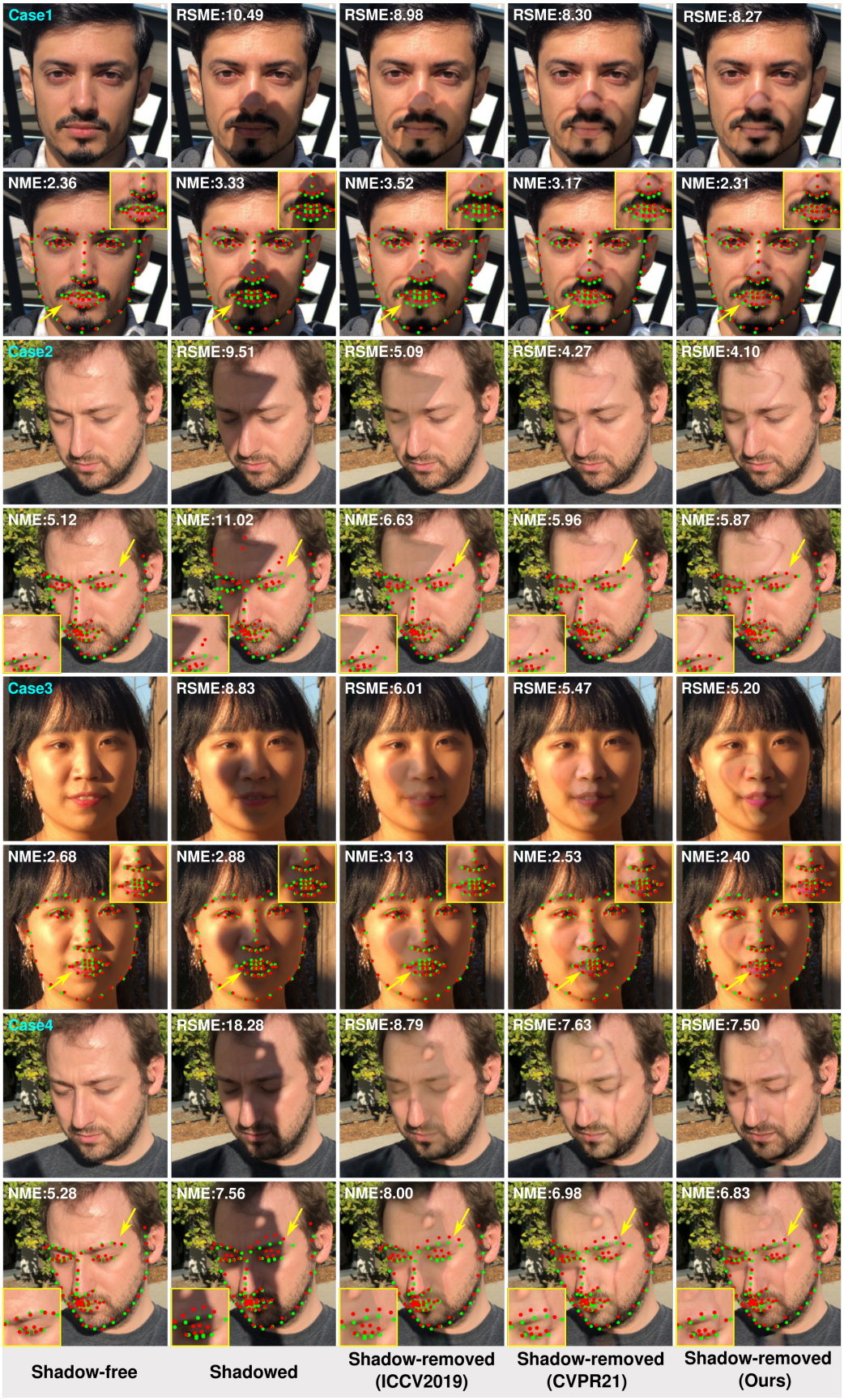

Figure 6.6: Shadow removal for facial landmark detection [22]. Red: prediction. Green: ground truth. RMSE measures the shadow removal accuracy, NME evaluates the detection performance. The lower, the better.

Visualization results We visualize the shadow removal and landmark detection performance of the proposed method compared to state-of-the-art shadow removal methods SP+M-Net (ICCV 2019) [69] and AEFNet (CVPR 2021) [26] on the real shadow dataset $\mathcal{D}_{\text{real}}$. We train the shadow removal methods with synthetic data $\mathcal{D}^t_{\text{syn}}$ and generalize such models to real data. As shown in Fig. 6.6, shadowed faces hurt the image quality with large root mean square error (RMSE), and presents unreasonable and much deteriorated landmark locations for the eyebrows (See Case2 and Case4) and mouth (See Case1 and Case3), as measured by much degraded NME scores. All the SP+M-Net, AEFNet and proposed method could obtain high visual-quality images with lower RMSE, as shown in Fig. 6.6, the landmark detection performance can almost be improved simultaneously. However, for the shadow removal with the landmark detection regularization (*i.e.*, the proposed method), the shadow-removed images achieve the best landmark detection performance and visual quality. For case1 and case3, the shadow-removed images achieves even better landmark detection performance compared to clean images. For example, the shadow-removed image by the proposed method achieves NME 2.40 compared to NME 2.68 of clean image for landmark detection evaluation for Case3. It turns out that the proposed method generalizes well to real data and experiments verify the effectiveness of the detection-aware shadow removal method.

6.6 Chapter Summary

In this chapter, we have proposed a shadow-removal benchmark dataset SHARELto explore the mutual influence of shadow removal and facial landmark detection tasks. We first proposed three strategies to construct the benchmark. Based on physical shadow model, we synthesize the shadowed faces considering four factors (*i.e.*, intensity, size, shape, and location) with three severities to cover diverse shadow patterns. We also proposed an adversarial shadow attack as hard shadow patterns to make the

landmark detection fail easily. Real shadowed face dataset for landmark detection is to reduce the distribution shift with synthetic data. Based on the proposed benchmark, we explored the shadow and shadow-removal effect on visual quality and landmark detection tasks comprehensively. We observed that there is a highly positive correlation between shadow removal and the facial landmark detection task, especially, when degradation level is higher. We then proposed a novel shadow-removal framework regularized by facial landmark detection to benefit each other. We verified the effectiveness of our proposed method in synthetic data, adversarial data and real data.

Chapter 7

Conclusion and future work

7.1 Conclusion

Adverse illumination is common and inevitable in the real-world applications, directly determining the image acquisition process. Image captured under adverse illumination suffers from image degradation, *e.g.*, shadow, underexposure, and noise. In this dissertation, we mainly study the problem of employing CNN-based image restoration for different applications, such as image visual quality enhancement, object detection improvement, and improving image visual quality and object detection at the same time.

In the research about shadow removal for image visual quality enhancement via auto-exposure fusion, we have proposed a novel and robust overexposure fusion method for performing shadow removal task. Multiple over-exposure, relighting each pixel with different exposures, could compensate each pixel individually to tackle position specified color and illumination degradation. It benefits the shadow removal task by recovering the natural image from the spatial variant color and illumination degradation. Shadow-aware FusionNet smartly fuses brackets of over-exposure shadow images with shadow image by an adaptive per-pixel kernel weight map. It helps to fully recover the background content preserving the color and illumination details. The proposed boundary-aware RefineNet further eliminates the remaining trace caused by the penumbra area along the shadow boundary. With the boundary loss added, by optimizing to preserve the non-shadow region and recover the ground-truth shadow-free area of the shadow image, our work can obtain traceless background with the state-of-the-art shadow removal performance on the ISTD, ISTD+, and SRD datasets.

In detail-preserving night-to-day image translation for nighttime vehicle detection improvement, we proposed a detail-preserving method to implement the nighttime to daytime image translation and thus adapt daytime trained detection model to nighttime detection. We firstly utilize style translation method to acquire paired

images of daytime and nighttime, which are hard to obtain in real-world applications. We propose to stylemix the reference styles to embody the diversity of synthetic nighttime scenarios. The following nighttime to daytime translation is implemented based on kernel prediction network to avoid texture corruption and trained with detection task to make the translated daytime image not only visually photo-realistic to the daytime scenario but also fit the detection task to reuse the daytime domain knowledge. The proposed method can perform both daytime and nighttime vehicle detection with one model. Experimental results showed that the proposed method achieved effective and accurate nighttime detection results.

In shadow removal for image visual quality enhancement and facial landmark detection improvement simultaneously, we have proposed a shadow-removal benchmark dataset SHAREL to explore the mutual influence of shadow removal and facial landmark detection tasks. We first proposed three strategies to construct the benchmark. Based on physical shadow model, we synthesize the shadowed faces considering four factors (*i.e.*, intensity, size, shape, and location) with three severities to cover diverse shadow patterns. We also proposed an adversarial shadow attack as hard shadow patterns to make the landmark detection fail easily. Real shadowed face dataset for landmark detection is to reduce the distribution shift with synthetic data. Based on the proposed benchmark, we explored the shadow and shadow-removal effects on visual quality and landmark detection tasks comprehensively. It turns out that there is a highly positive correlation between shadow removal and the facial landmark detection task, especially, when degradation level is higher. We then proposed a novel shadow-removal framework regularized by facial landmark detection to further improve their performance. We verified the effectiveness of our proposed method in synthetic data, adversarial data and real data.

7.2 Future work

Based on the above study on image restoration under adverse illumination for different applications, we can outlook some of the future works. Image restoration under adverse illumination is important and worth devoting effort to study, since adverse illumination is widely existing in image collection process. Besides the topic we studied in this dissertation, there still exist a lot of problems to be addressed for image restoration under adverse illumination.

Nighttime object detection. In the autonomous driving application, knowledge transfer from daytime perception model to nighttime scenario is an indispensable and challenging part. Image restoration or image translation between daytime domain and nighttime domain can improve the nighttime image visual quality directly. An important research topic is to explore image restoration for recovering the degraded features for object detection and semantic segmentation. There are many different object classes in autonomous driving, such as pedestrians, cross walks, cars, and buildings, which is more challenging and difficult to perceive compared to vehicle detection in the traffic surveillance discussed in this dissertation.

Image restoration under various illumination levels. Image restoration models, considering more illumination levels, such as weak lighting, underexposure, moonlight, twilight, dim light, dark, extremely dark, backlit, non-uniform light, and colored light, are worth to be explored to construct a powerful CNN for empowering more strong generalization ability to achieve better visually pleasing images. In this dissertation, we only perform shadow removal to improve the image visual quality, and the illumination variation is very limited in the shadow patterns. In real-world applications, due to complicated environment, illumination levels are diverse.

Effective network design and learning strategy. Current state-of-the-art image restoration methods under adverse illumination are CNN or GAN based networks. Recently, Transformer-based networks have achieved great progress in a large amount of computer vision tasks [7, 88, 120, 23] due to its self-attention mechanism to capture global interactions between contexts. Vision Transformers for image restoration under adverse illumination is a promising direction to explore. In addition, current deep learning based image restoration methods mainly utilize supervised learning that requires massive paired training data, which may overfit on a specific dataset. Recently, unsupervised learning [53] and zero-shot learning [36] have shown robust performance for real scenes without paired training data, which is easy to be applied to real-world applications.

Bibliography

[1] José M Álvarez Alvarez and Antonio M Ĺopez, *Road detection based on illuminant invariance*, IEEE Transactions on Intelligent Transportation Systems **12** (2010), no. 1, 184–193.

[2] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool, *Night-to-day image translation for retrieval-based localization*, IEEE International Conference on Robotics and Automation, 2019, pp. 5958–5964.

[3] Vasileios Belagiannis and Andrew Zisserman, *Recurrent human pose estimation*, International Conference on Automatic Face & Gesture Recognition, 2017, pp. 468–475.

[4] Chinmay Belthangady and Loic A Royer, *Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction*, Nature methods **16** (2019), no. 12, 1215–1225.

[5] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou, *Large scale 3d morphable models*, International Journal of Computer Vision **126** (2018), no. 2, 233–254.

[6] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang, *Toward real-world single image super-resolution: A new benchmark and a new model*, IEEE International Conference on Computer Vision, 2019, pp. 3086–3095.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, *End-to-end object detection with transformers*, European Conference on Computer Vision, 2020, pp. 213–229.

[8] Samprit Chatterjee, Ali S Hadi, et al., *Influential observations, high leverage points, and outliers in linear regression*, Statistical science **1** (1986), no. 3, 379–393.

[9] Liang Chen, Jinshan Pan, Junjun Jiang, Jiawei Zhang, Zhen Han, and Linchao Bao, *Multi-stage degradation homogenization for super-resolution of face images*

with extreme degradations, IEEE Transactions on Image Processing **30** (2021), 5600–5612.

[10] Xinqiang Chen, Zhibin Li, Yongsheng Yang, Lei Qi, and Ruimin Ke, *High-resolution vehicle trajectory extraction and denoising from aerial videos*, IEEE Transactions on Intelligent Transportation Systems (2020).

[11] Yinpeng Chen and Hari Sundaram, *Estimating complexity of 2d shapes*, Workshop on Multimedia Signal Processing, 2005, pp. 1–4.

[12] Yupeng Cheng, Felix Juefei-Xu, Qing Guo, Huazhu Fu, Xiaofei Xie, Shang-Wei Lin, Weisi Lin, and Yang Liu, *Adversarial exposure attack on diabetic retinopathy imagery*, arXiv preprint arXiv:2009.09231 (2020).

[13] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon, *Kaist multi-spectral day/night data set for autonomous and assisted driving*, IEEE Transactions on Intelligent Transportation Systems **19** (2018), no. 3, 934–948.

[14] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor, *Active appearance models*, IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001), no. 6, 681–685.

[15] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham, *Active shape models-their training and application*, Computer Vision and Image Understanding **61** (1995), no. 1, 38–59.

[16] David Cristinacce and Timothy F Cootes, *Feature detection and tracking with constrained local models.*, BMVC, 2006.

[17] Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati, *Detecting moving objects, ghosts, and shadows in video streams*, IEEE Transactions on Pattern Analysis and Machine Intelligence **25** (2003), no. 10, 1337–1342.

[18] Xiaodong Cun, Chi-Man Pun, and Cheng Shi, *Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan.*, AAAI Conference on Artificial Intelligence, 2020, pp. 10680–10687.

[19] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang, *Second-order attention network for single image super-resolution*, IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11065–11074.

[20] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao, *Argan: Attentive recurrent generative adversarial network for shadow detection and removal*, IEEE International Conference on Computer Vision, 2019, pp. 10213–10222.

[21] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, *Learning a deep convolutional network for image super-resolution*, European Conference on Computer Vision, 2014, pp. 184–199.

[22] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang, *Style aggregated network for facial landmark detection*, IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 379–388.

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

[24] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao, *Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping*, IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2427–2436.

[25] Lan Fu, Hongkai Yu, Felix Juefei-Xu, Jinlong Li, Qing Guo, and Song Wang, *Let there be light: Improved traffic surveillance via detail preserving night-to-day transfer*, IEEE Transactions on Circuits and Systems for Video Technology (2021).

[26] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang, *Auto-exposure fusion for single-image shadow removal*, IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 10571–10580.

[27] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding, *A weighted variational model for simultaneous reflectance and illumination estimation*, IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2782–2790.

[28] Ruijun Gao, Qing Guo, Felix Juefei-Xu, Hongkai Yu, and Wei Feng, *AdvHaze: Adversarial Haze Attack*, arXiv preprint arXiv:2104.13673 (2021).

[29] Ruijun Gao, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Huazhu Fu, Wei Feng, Yang Liu, and Song Wang, *Can you spot the chameleon? adversarially camou-*

flaging images from co-salient object detection, IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 2150–2159.

[30] Ruijun Gao, Qing Guo, Qian Zhang, Felix Juefei-Xu, Hongkai Yu, and Wei Feng, *Adversarial Relighting against Face Recognition*, arXiv preprint arXiv:2108.07920 (2021).

[31] Junfeng Ge, Yupin Luo, and Gyomei Tei, *Real-time pedestrian detection and tracking at nighttime for driver-assistance systems*, IEEE Transactions on Intelligent Transportation Systems **10** (2009), no. 2, 283–298.

[32] Ross Girshick, *Fast r-cnn*, IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

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

[34] Han Gong and Darren Cosker, *Interactive removal and ground truth for difficult shadow scenes*, JOSA A **33** (2016), no. 9, 1798–1811.

[35] Maciej Gryka, Michael Terry, and Gabriel J Brostow, *Learning to remove soft shadows*, ACM Transactions on Graphics **34** (2015), no. 5, 1–15.

[36] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong, *Zero-reference deep curve estimation for low-light image enhancement*, IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 1780–1789.

[37] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li, *Towards fast, accurate and stable 3d dense face alignment*, European Conference on Computer Vision, 2020, pp. 152–168.

[38] Qing Guo, Ziyi Cheng, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yang Liu, and Jianjun Zhao, *Learning to Adversarially Blur Visual Object Tracking*, IEEE International Conference on Computer Vision, October 2021.

[39] Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Jian Wang, Bing Yu, Wei Feng, and Yang Liu, *Watch out! Motion is Blurring the Vision of Your Deep Neural Networks*, Advances in Neural Information Processing Systems, 2020.

[40] Qing Guo, Jingyang Sun, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Wei Feng, and Yang Liu, *Efficientderain: Learning pixel-wise dilation filtering for high-efficiency single-image deraining*, AAAI Conference on Artificial Intelligence, 2021.

[41] Qing Guo, Xiaofei Xie, Felix Juefei-Xu, Lei Ma, Zhongguo Li, Wanli Xue, Wei Feng, and Yang Liu, *SPARK: Spatial-aware Online Incremental Attack Against Visual Tracking*, European Conference on Computer Vision, Aug 2020.

[42] Ruiqi Guo, Qieyun Dai, and Derek Hoiem, *Paired regions for shadow detection and removal*, IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2012), no. 12, 2956–2967.

[43] Xiaojie Guo, Yu Li, and Haibin Ling, *Lime: Low-light image enhancement via illumination map estimation*, IEEE Transactions on Image Processing **26** (2016), no. 2, 982–993.

[44] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan, *Closed-loop matters: Dual regression networks for single image super-resolution*, IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 5407–5416.

[45] Junwei Han, Le Yang, Dingwen Zhang, Xiaojun Chang, and Xiaodan Liang, *Reinforcement cutting-agent learning for video object segmentation*, IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9080–9089.

[46] Junwei Han, Xiwen Yao, Gong Cheng, Xiaoxu Feng, and Dong Xu, *P-cnn: Part-based convolutional neural networks for fine-grained visual categorization*, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).

[47] Pat Hanrahan and Wolfgang Krueger, *Reflection from layered surfaces due to subsurface scattering*, Conference on Computer Graphics and Interactive Techniques, 1993, pp. 165–174.

[48] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, *Mask r-cnn*, IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.

[49] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan, *Augmix: A simple data processing method to improve robustness and uncertainty*, International Conference on Learning Representations, 2019.

[50] Mazin Hnewa and Hayder Radha, *Object detection under rainy conditions for autonomous vehicles: A review of state-of-the-art and emerging techniques*, IEEE Signal Processing Magazine **38** (2020), no. 1, 53–67.

[51] Chang-Hui Hu, Jian Yu, Fei Wu, Yang Zhang, Xiao-Yuan Jing, Xiao-Bo Lu, and Pan Liu, *Face illumination recovery for the deep learning feature under severe illumination variations*, Pattern Recognition **111** (2021), 107724.

[52] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng, *Direction-aware spatial context features for shadow detection and removal*, IEEE Transactions on Pattern Analysis and Machine Intelligence **42** (2020), no. 11, 2795–2808.

[53] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng, *Mask-shadowgan: Learning to remove shadows from unpaired data*, IEEE International Conference on Computer Vision, 2019, pp. 2472–2481.

[54] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, *Multimodal unsupervised image-to-image translation*, European Conference on Computer Vision, 2018, pp. 172–189.

[55] Yihao Huang, Felix Juefei-Xu, Qing Guo, Weikai Miao, Yang Liu, and Geguang Pu, *Advbokeh: Learning to adversarially defocus blur*, arXiv preprint (2021).

[56] Naoto Inoue and Toshihiko Yamasaki, *Learning from synthetic shadows for shadow detection and removal*, IEEE Transactions on Circuits and Systems for Video Technology (2020).

[57] Sergey Ioffe and Christian Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, International Conference on Machine Learning, 2015, pp. 448–456.

[58] Junjun Jiang, Jiayi Ma, Chen Chen, Xinwei Jiang, and Zheng Wang, *Noise robust face image super-resolution through smooth sparse representation*, IEEE Transactions on Cybernetics **47** (2016), no. 11, 3991–4002.

[59] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang, *Enlightengan: Deep light enhancement without paired supervision*, IEEE Transactions on Image Processing **30** (2021), 2340–2349.

[60] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, *Perceptual losses for real-time style transfer and super-resolution*, European Conference on Computer Vision, 2016, pp. 694–711.

[61] Felix Juefei-Xu and Marios Savvides, *An image statistics approach towards efficient and robust refinement for landmarks on facial boundary*, International Conference on Biometrics: Theory, Applications and Systems, 2013, pp. 1–8.

[62] Cláudio Rosito Jung, *Efficient background subtraction and shadow removal for monochromatic video sequences*, IEEE Transactions on Multimedia **11** (2009), no. 3, 571–577.

[63] Ruimin Ke, Zhibin Li, Jinjun Tang, Zewen Pan, and Yinhai Wang, *Real-time traffic flow parameter estimation from uav video based on ensemble classifier and optical flow*, IEEE Transactions on Intelligent Transportation Systems **20** (2018), no. 1, 54–64.

[64] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, *Accurate image super-resolution using very deep convolutional networks*, IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 1646–1654.

[65] Hulin Kuang, Kai-Fu Yang, Long Chen, Yong-Jie Li, Leanne Lai Hang Chan, and Hong Yan, *Bayes saliency-based object proposal generator for nighttime traffic images*, IEEE Transactions on Intelligent Transportation Systems **19** (2017), no. 3, 814–825.

[66] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng, *Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood*, IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 8236–8246.

[67] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang, *Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better*, IEEE International Conference on Computer Vision, 2019, pp. 8878–8887.

[68] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, *Deep laplacian pyramid networks for fast and accurate super-resolution*, IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 624–632.

[69] Hieu Le and Dimitris Samaras, *Shadow removal via shadow image decomposition*, IEEE International Conference on Computer Vision, 2019, pp. 8578–8587.

[70] Hieu Le and Dimitris Samaras, *From shadow segmentation to shadow removal*, European Conference on Computer Vision, 2020.

[71] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao, *An underwater image enhancement benchmark dataset and beyond*, IEEE Transactions on Image Processing **29** (2019), 4376–4389.

[72] Jie Li, Katherine A Skinner, Ryan M Eustice, and Matthew Johnson-Roberson, *Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images*, IEEE Robotics and Automation letters **3** (2017), no. 1, 387–394.

[73] Jinlong Li, Zhigang Xu, Lan Fu, Xuesong Zhou, and Hongkai Yu, *Domain adaptation from daytime to nighttime: A situation-sensitive vehicle detection and traffic flow parameter estimation framework*, Transportation Research Part C: Emerging Technologies **124** (2021), 102946.

[74] Runde Li, Jinshan Pan, Zechao Li, and Jinhui Tang, *Single image dehazing via conditional generative adversarial network*, IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8202–8211.

[75] Shuguang Li, Hongkai Yu, Jingru Zhang, Kaixin Yang, and Ran Bin, *Video-based traffic data collection system for multiple vehicle types*, IET Intelligent Transport Systems **8** (2014), no. 2, 164–174.

[76] Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, et al., *Structured landmark detection via topology-adapting deep graph learning*, European Conference on Computer Vision, 2020, pp. 266–283.

[77] Yiming Li, Congcong Wen, Felix Juefei-Xu, and Chen Feng, *Fooling LiDAR Perception via Adversarial Trajectory Perturbation*, IEEE International Conference on Computer Vision, October 2021.

[78] Yu Li, Sheng Tang, Rui Zhang, Yongdong Zhang, Jintao Li, and Shuicheng Yan, *Asymmetric gan for unpaired image-to-image translation*, IEEE Transactions on Image Processing **28** (2019), no. 12, 5881–5896.

[79] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu, *Celeb-df: A large-scale challenging dataset for deepfake forensics*, IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207–3216.

[80] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, *Focal loss for dense object detection*, IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[81] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang, *Non-local recurrent network for image restoration*, Advances in Neural Information Processing systems **31** (2018).

[82] Ding Liu, Bihan Wen, Xianming Liu, Zhangyang Wang, and Thomas S Huang, *When image denoising meets high-level vision tasks: A deep learning approach*, International Joint Conference on Artificial Intelligence, 2017, pp. 842–848.

[83] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu, *Joint face alignment and 3d face reconstruction*, European Conference on Computer Vision, 2016, pp. 545–560.

[84] Ming-Yu Liu, Thomas Breuel, and Jan Kautz, *Unsupervised image-to-image translation networks*, Advances in Neural Information Processing Systems, 2017, pp. 700–708.

[85] Nian Liu, Ni Zhang, and Junwei Han, *Learning selective self-mutual attention for rgb-d saliency detection*, IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 13756–13765.

[86] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, *Ssd: Single shot multibox detector*, European Conference on Computer Vision, 2016, pp. 21–37.

[87] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang, *Exploring disentangled feature representation beyond face identification*, IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2080–2089.

[88] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, *Swin transformer: Hierarchical vision transformer using shifted windows*, IEEE International Conference on Computer Vision, 2021, pp. 10012–10022.

[89] Yiqun Mei, Yuchen Fan, and Yuqian Zhou, *Image super-resolution with non-local sparse attention*, IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 3517–3526.

[90] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll, *Burst denoising with kernel prediction networks*, IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2502–2510.

[91] Hajime Nada, Vishwanath A Sindagi, He Zhang, and Vishal M Patel, *Pushing the limits of unconstrained face detection: a challenge dataset and baseline results*, International Conference on Biometrics Theory, Applications and Systems, 2018, pp. 1–10.

[92] Sohail Nadimi and Bir Bhanu, *Physical models for moving shadow and object detection in video*, IEEE Transactions on Pattern Analysis and Machine Intelligence **26** (2004), no. 8, 1079–1087.

[93] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu, *Contrastive learning for unpaired image-to-image translation*, European Conference on Computer Vision, Springer, 2020, pp. 319–345.

[94] Yanting Pei, Yaping Huang, Qi Zou, Xingyuan Zhang, and Song Wang, *Effects of image degradation and degradation removal to cnn-based image classification*, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).

[95] Patrick Pérez, Michel Gangnet, and Andrew Blake, *Poisson image editing*, ACM SIGGRAPH (2003), 313–318.

[96] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau, *Deshadownet: A multi-context embedding deep network for shadow removal*, IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4067–4075.

[97] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, *You only look once: Unified, real-time object detection*, IEEE International Conference on Computer Vision, 2016, pp. 779–788.

[98] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, *Faster r-cnn: Towards real-time object detection with region proposal networks*, Advances in Neural Information Processing Systems, 2015, pp. 91–99.

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

[100] German Ros and Jose M Alvarez, *Unsupervised image transformation for outdoor semantic labelling*, IEEE Intelligent Vehicles Symposium, 2015, pp. 537–542.

[101] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, *300 faces in-the-wild challenge: The first facial landmark localization challenge*, IEEE International Conference on Computer Vision Workshops, 2013, pp. 397–403.

[102] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin, *Static and dynamic 3d facial expression recognition: A comprehensive survey*, Image and Vision Computing **30** (2012), no. 10, 683–697.

[103] Andres Sanin, Conrad Sanderson, and Brian C Lovell, *Improved shadow removal for robust person tracking in surveillance scenarios*, International Conference on Pattern Recognition, 2010, pp. 141–144.

[104] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn, *Deformable model fitting by regularized landmark mean-shift*, International Journal of Computer Vision **91** (2011), no. 2, 200–215.

[105] Ravi Kumar Satzoda and Mohan Manubhai Trivedi, *Looking at vehicles in the night: Detection and dynamics of rear lights*, IEEE Transactions on Intelligent Transportation Systems (2016).

[106] Yael Shor and Dani Lischinski, *The shadow meets the mask: Pyramid-based shadow removal*, Computer Graphics Forum, vol. 27, 2008, pp. 577–586.

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

[108] Zhun Sun, Mete Ozay, Yan Zhang, Xing Liu, and Takayuki Okatani, *Feature quantization for defending against distortion of images*, IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7957–7966.

[109] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu, *Memnet: A persistent memory network for image restoration*, IEEE International Conference on Computer Vision, 2017, pp. 4539–4547.

[110] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, *Face2face: Real-time face capture and reenactment of rgb*

videos, IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2387–2395.

[111] Binyu Tian, Qing Guo, Felix Juefei-Xu, Wen Le Chan, Yupeng Cheng, Xiaohong Li, Xiaofei Xie, and Shengchao Qin, *Bias Field Poses a Threat to DNN-Based X-Ray Recognition*, IEEE International Conference on Multimedia and Expo, 2021.

[112] Binyu Tian, Felix Juefei-Xu, Qing Guo, Xiaofei Xie, Xiaohong Li, and Yang Liu, *AVA: Adversarial Vignetting Attack against Visual Recognition*, International Joint Conference on Artificial Intelligence, 2021.

[113] Alexander Toshev and Christian Szegedy, *Deeppose: Human pose estimation via deep neural networks*, IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653–1660.

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

[115] Roberto Valle, Jose M Buenaposada, Antonio Valdes, and Luis Baumela, *A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment*, European Conference on Computer Vision, 2018, pp. 585–601.

[116] Tomas F Yago Vicente, Minh Hoai, and Dimitris Samaras, *Leave-one-out kernel optimization for shadow detection and removal*, IEEE Transactions on Pattern Analysis and Machine Intelligence **40** (2017), no. 3, 682–695.

[117] Jifeng Wang, Xiang Li, and Jian Yang, *Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal*, IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1788–1797.

[118] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al., *Deep high-resolution representation learning for visual recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[119] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo, *Learning parallax attention for stereo image super-resolution*, IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12250–12259.

[120] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li, *Transformer meets tracker: Exploiting temporal context for robust visual tracking*, IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 1571–1580.

[121] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan, *Learning from synthetic data for crowd counting in the wild*, IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8198–8207.

[122] Run Wang, Felix Juefei-Xu, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Yang Liu, *Amora: Black-box Adversarial Morphing Attack*, ACM International Conference on Multimedia, 2020.

[123] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li, *Naturalness preserved enhancement algorithm for non-uniform illumination images*, IEEE Transactions on Image Processing **22** (2013), no. 9, 3538–3548.

[124] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, *Non-local neural networks*, IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[125] Yunxuan Wei, Shuhang Gu, Yawei Li, Radu Timofte, Longcun Jin, and Hengjie Song, *Unsupervised real-world image super resolution via domain-distance aware training*, IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 13385–13394.

[126] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou, *Look at boundary: A boundary-aware face alignment algorithm*, IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2129–2138.

[127] Yue Wu, Chao Gou, and Qiang Ji, *Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion*, IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 3471–3480.

[128] Yue Wu and Qiang Ji, *Facial landmark detection: A literature survey*, International Journal of Computer Vision **127** (2019), no. 2, 115–142.

[129] Chunxia Xiao, Ruiyun She, Donglin Xiao, and Kwan-Liu Ma, *Fast shadow removal using adaptive multi-scale illumination transfer*, Computer Graphics Forum, vol. 32, 2013, pp. 207–218.

[130] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, *Aggregated residual transformations for deep neural networks*, IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1492–1500.

[131] Fengliang Xu, Xia Liu, and Kikuo Fujimura, *Pedestrian detection and tracking with night vision*, IEEE Transactions on Intelligent Transportation Systems **6** (2005), no. 1, 63–71.

[132] Le Yang, Junwei Han, Dingwen Zhang, Nian Liu, and Dong Zhang, *Segmentation in weakly labeled videos via a semantic ranking and optical warping network*, IEEE Transactions on Image Processing **27** (2018), no. 8, 4025–4037.

[133] Xiwen Yao, Junwei Han, Gong Cheng, Xueming Qian, and Lei Guo, *Semantic annotation of high-resolution satellite images via weakly supervised learning*, IEEE Transactions on Geoscience and Remote Sensing (2016), 3660–3671.

[134] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell, *Bdd100k: A diverse driving dataset for heterogeneous multitask learning*, IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 2636–2645.

[135] Hongkai Yu, Dazhou Guo, Zhipeng Yan, Lan Fu, Jeff Simmons, Craig P Przybyla, and Song Wang, *Weakly supervised easy-to-hard learning for object detection in image sequences*, Neurocomputing **398** (2020), 71–82.

[136] Liming Zhai, Felix Juefei-Xu, Qing Guo, Xiaofei Xie, Lei Ma, Wei Feng, Shengchao Qin, and Yang Liu, *It's raining cats or dogs? adversarial rain attack on dnn perception*, arXiv preprint arXiv:2009.09205 (2020).

[137] Dingwen Zhang, Guangyu Guo, Dong Huang, and Junwei Han, *Poseflow: A deep motion representation for understanding human behaviors in videos*, IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6762–6770.

[138] Dingwen Zhang, Junwei Han, Le Yang, and Dong Xu, *Spftn: a joint learning framework for localizing and segmenting objects in weakly labeled videos*, IEEE Transactions on Pattern Analysis and Machine Intelligence (2018).

[139] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan, *Freenet: Multi-identity face reenactment*, IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 5326–5335.

[140] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte, *Plug-and-play image restoration with deep denoiser prior*, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[141] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, *Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising*, IEEE Transactions on Image Processing **26** (2017), no. 7, 3142–3155.

[142] Kai Zhang, Wangmeng Zuo, and Lei Zhang, *Ffdnet: Toward a fast and flexible solution for cnn-based image denoising*, IEEE Transactions on Image Processing **27** (2018), no. 9, 4608–4622.

[143] Ling Zhang, Qing Zhang, and Chunxia Xiao, *Shadow remover: Image shadow removal based on illumination recovering optimization*, IEEE Transactions on Image Processing **24** (2015), no. 11, 4623–4636.

[144] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen, *Improving shadow suppression for illumination robust face recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence **41** (2018), no. 3, 611–624.

[145] Xuaner Zhang, Jonathan T Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E Jacobs, *Portrait shadow manipulation*, ACM Transactions on Graphics **39** (2020), no. 4, 78–1.

[146] Yujia Zhang, Xiaodan Liang, Dingwen Zhang, Min Tan, and Eric P Xing, *Unsupervised object-level video summarization with online motion auto-encoder*, Pattern Recognition Letters **130** (2020), 376–385.

[147] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, *Image super-resolution using very deep residual channel attention networks*, European Conference on Computer Vision, 2018, pp. 286–301.

[148] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, *Residual dense network for image super-resolution*, IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481.

[149] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, *Facial landmark detection by deep multi-task learning*, European Conference on Computer Vision, Springer, 2014, pp. 94–108.

[150] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia, *Learning self-consistency for deepfake detection*, IEEE International Conference on Computer Vision, 2021, pp. 15023–15033.

[151] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, *Unpaired image-to-image translation using cycle-consistent adversarial networks*, IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[152] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng, *Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection*, European Conference on Computer Vision, 2018, pp. 121–136.

[153] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li, *High-fidelity pose and expression normalization for face recognition in the wild*, IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 787–796.

[154] Xu Zou, Sheng Zhong, Luxin Yan, Xiangyun Zhao, Jiahuan Zhou, and Ying Wu, *Learning robust facial landmark detection via hierarchical structured ensemble*, IEEE International Conference on Computer Vision, 2019, pp. 141–150.