Develop then Rival: A Human Vision-Inspired Framework for Superimposed Image Decomposition

Huiyu Duan, Wei Shen, Xiongkuo Min, Yuan Tian, Jae-Hyun Jung, Xiaokang Yang, *Fellow, IEEE*, and Guangtao Zhai, *Senior Member, IEEE*

***Abstract*—A single superimposed image containing two image views causes visual confusion for both human vision and computer vision. Human vision needs a "develop-then-rival" process to decompose the superimposed image into two individual images, which effectively suppresses visual confusion. However, separating individual image views from a single superimposed image has been an important but challenging task in computer vision area for a long time. In this paper, we propose a human vision-inspired framework for single superimposed image decomposition. We first propose a network to simulate the development stage, which tries to understand and distinguish the semantic information of the two layers of a single superimposed image. To further simulate the rivalry activation/suppression process in human brains, we carefully design a rivalry stage, which incorporates the original mixed input (superimposed image), the activated visual information (outputs of the development stage) together, and then rivals to get images without ambiguity. Experimental results show that our novel framework effectively separates the superimposed images and significantly improves the performance with better output quality compared with state-of-the-art methods. The proposed method also achieves state-of-the-art results on related applications including single image reflection removal, single image rain removal, single image shadow removal, and illumination correction, *etc.*, which validates the generalization of the framework.**

***Index Terms*—Superimposed image decomposition, develop then rival, reflection removal, rain removal, shadow removal, illumination correction.**

I. Introduction

Visual confusion [1], [2] (the perceptions of two different views are superimposed onto the same space) is frequently encountered when viewing a single superimposed image and may arise the ambiguity for both human vision and computer vision. Thus, the topics related to separating superimposed images including reflection removal [3], [4], image deraining [5], [6], shadow removal [7], [8], image dehazing [9], *etc.*, have long been important tasks in computer vision field, which aim at not only generating high-quality images in accordance with human vision, but also benefiting the downstream computer vision tasks, *e.g.,* image classification, object detection, *etc* [10]. Let I be the input image with superimposed layers, it can be approximately modeled as a combination of two image layers I_1 and I_2, *i.e.*, $I = g(I_1) + f(I_2)$, where $g(\cdot)$ and $f(\cdot)$ denote various degradations for I_1 and I_2, respectively. When only given a single input image I, there are an infinite number of feasible decompositions to recover I_1 and I_2. Therefore, separating a single superimposed image is an ill-posed problem [11], not only due to the unknown mixing function, but also because of the lack of constraints on the output space.

Previous statistics-based superimposed image separation methods have been studied for a long time [11]–[14]. However, these methods need heavy user interactions or require a series of multiple mixed inputs. Recently, deep learning-based approaches have been extensively studied on image decomposition related applications and made great progress [5], [7], [9], [15]–[18]. Nevertheless, most of them only focused on one specific separation case, while a unified framework is rarely considered. Gandelsman *et al.* [19] have proposed a unified framework named "Double-DIP" for unsupervised image decomposition. Although this method can well handle the input with regular mixed patterns, they struggle with the decomposition of natural images. Zou *et al.* [20] have proposed a unified framework for supervised image decomposition based on Generative Adversarial Network (GAN). However, the separated two images still have residual information from each other.

Human vision utilizes monocular rivalry to eliminate the ambiguity caused by visual confusion. For a single superimposed image, human vision usually takes a while to develop monocular rivalry (Stage I), then alternatively activates/suppresses one image layer [21]–[23] to eliminate visual confusion during monocular rivalry (Stage II). For example, when looking through a transparent glass, a transmission scene and a reflection scene can be seen simultaneously. Humans first need a while to understand and distinguish the semantic

Manuscript received July 16, 2021; revised March 5, 2022; accepted April 20, 2022. Date of publication XXX XX, XXXX; date of current version XXX XX, XXXX. This work was supported in part by the National Natural Science Foundation of China under Grant 61831015, Grant 61901260, and Grant 62176159, in part by the National Key R&D Program of China 2021YFE0206700, in part by the Natural Science Foundation of Shanghai 21ZR1432200, and in part by the Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Juyong Zhang. *(Corresponding authors: Wei Shen; Guangtao Zhai.)*

Huiyu Duan, Wei Shen, Xiongkuo Min, Yuan Tian, Xiaokang Yang, and Guangtao Zhai are with MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, 200240, China (e-mail: huiyuduan@sjtu.edu.cn; shenwei1231@gmail.com; minxiongkuo@sjtu.edu.cn; ee_tianyuan@sjtu.edu.cn; xkyang@sjtu.edu.cn; zhaiguangtao@sjtu.edu.cn).

Jae-Hyun Jung is with the Schepens Eye Research Institute, Harvard Medical School, Boston, 02114, USA (e-mail: Jae-hyun_Jung@MEEI.HARVARD.EDU).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier XXXXXX

information of the transmission layer and the reflection layer, respectively. Then the attentions on two layers compete with each other to form monocular rivalry, which causes that during one period, only one layer is activated and another layer is suppressed. In this work, inspired by this *develop-then-rival* process of human vision, we propose a unified framework for single superimposed image separation, which also consists of a development stage and a rivalry stage.

Similar to human vision, the first part of our framework, termed a *development stage*, tries to understand the features of the superimposed image and then roughly classifies them into two layers. Since the main network in the development stage requires a strong feature learning ability to better disentangle superimposed features, we first improve the basic network [20] (a U-Net) to a *differentiation net* by incorporating the contextual attention information [24], [25] for better feature learning, as well as adding dilated convolutional layers and non-local layers to original convolutional layers for enlarging receptive field. Then, we introduce a multi-scale [26] crossroad perceptual loss, which *crossly* compares the feature difference between the outputs of multi-scale deconvolutional layers and ground truths, thereby enforcing each deconvolutional layer to learn the task related features.

The second part of our framework, termed a *rivalry stage*, simulates the activation/suppression process of monocular rivalry of human vision, and tries to activate one superimposed layer and suppress another superimposed layer through a dual-pathway network. Since the sequence of the two predicted images obtained from the development stage may not match the sequence of the two ground truths, in this stage, we introduce a "crossroad judgement", which judges the sequence and matches one activated prediction image (from the development stage) to its target ground truth. Next we take the original superimposed image and the activated layer as inputs, and then use an *activation net* to enhance the activated layer and suppress another layer. Finally, to further improve the perceptual quality of the outputs and avoid confusion, we propose a confusion loss which restricts the residual information left by the non-target images.

The proposed framework also follows the coarse-to-fine generative process. Specifically, it first coarsely decomposes the superimposed images into two parts, then further leverages this prior information to activate the selected layer in the superimposed image and refines to get a higher-quality image layer. Compared to [20], our contextual attention module and multi-scale crossroad perceptual loss can better separate the superimposed images, and the proposed second network can effectively remove the residual information of the two separated images. In summary, our main contributions are:

- We propose a novel superimposed image decomposition framework inspired by the "develop-then-rival" process of human vision, which consists of a development stage and a rivalry stage.
- In the development stage, we leverage the contextual attention information within channels for better feature learning, and propose the multi-scale crossroad perceptual loss to enable the framework to learn the prior knowledge of the decomposition as early as possible.
- In the rivalry stage, we propose a strategy to simulate the activation/suppression process of monocular rivalry, and introduce a confusion loss to suppress the information from the non-target ground truth.
- Extensive experiments show that the proposed model achieves state-of-the-art results on the superimposed image separation task, as well as related applications, including single image reflection removal, single image rain removal, single image shadow removal, and illumination correction, *etc*.

II. Related Work

Visual confusion. The superimposition of two views of the visual scene (*i.e.*, two images in this paper) allows people to see two different things in one direction, which may result in visual confusion [27] and influence the perceptual quality [2], [28]–[32]. In this paper, we only consider monocular visual confusion (visual confusion within one eye), which may lead to monocular rivalry [1], [21]. Monocular rivalry needs a while to develop [21], [22] and it possibly occurs only with attention competition [1]. O'Shea *et al.* [22] have presented that the alternative of monocular rivalry is more dependent on semantic attention processes. Therefore, applying contextual attention for extracted features may improve the ability of the decomposition network. Moreover, the peripheral area of the eye and the eye movement enable human to perceive the texture information from the non-fixation areas [33]. It is also important to consider enlarging the receptive field of the network.

Superimposed image separation. In the field of signal processing, the separation of several individual signals from the mixed input has been studied for a long time, which is known as Blind Source Separation (BSS) [34]. Hyvarinen *et al.* have proposed the Independent Component Analysis (ICA) for BSS, which is based on the theorem that a sum of independent variables tends to be a Gaussian distribution under certain conditions. Based on the ICA, some statistics-based methods have been proposed to measure the independence and non-Gaussianity of the superimposed images for separating individual layers from them, which require a series of multiple mixed inputs [12]–[14] or additional user interactions [11], while these additional parameters are not always available in practice. Recently, an unsupervised method for superimposed image separation named "Double-DIP" has been proposed [19]. However, this work can handle the input with mixed patterns, but may not perform well for natural image superimposition separation. Zou *et al.* [20] have proposed a supervised deep adversarial decomposition method (denoted as DAD below) based on GAN. Nevertheless, the separated two images still have residual information from each other and the refinement for the obtained separated images is lacking.

Various separation tasks. Many computer vision tasks can be expressed as the superimposed image decomposition problem:

1) Single image reflection removal. An image with reflection can be seen as a transmission layer image superimposed by a reflection image. Some traditional methods of reflection

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3172882, IEEE Transactions on Multimedia

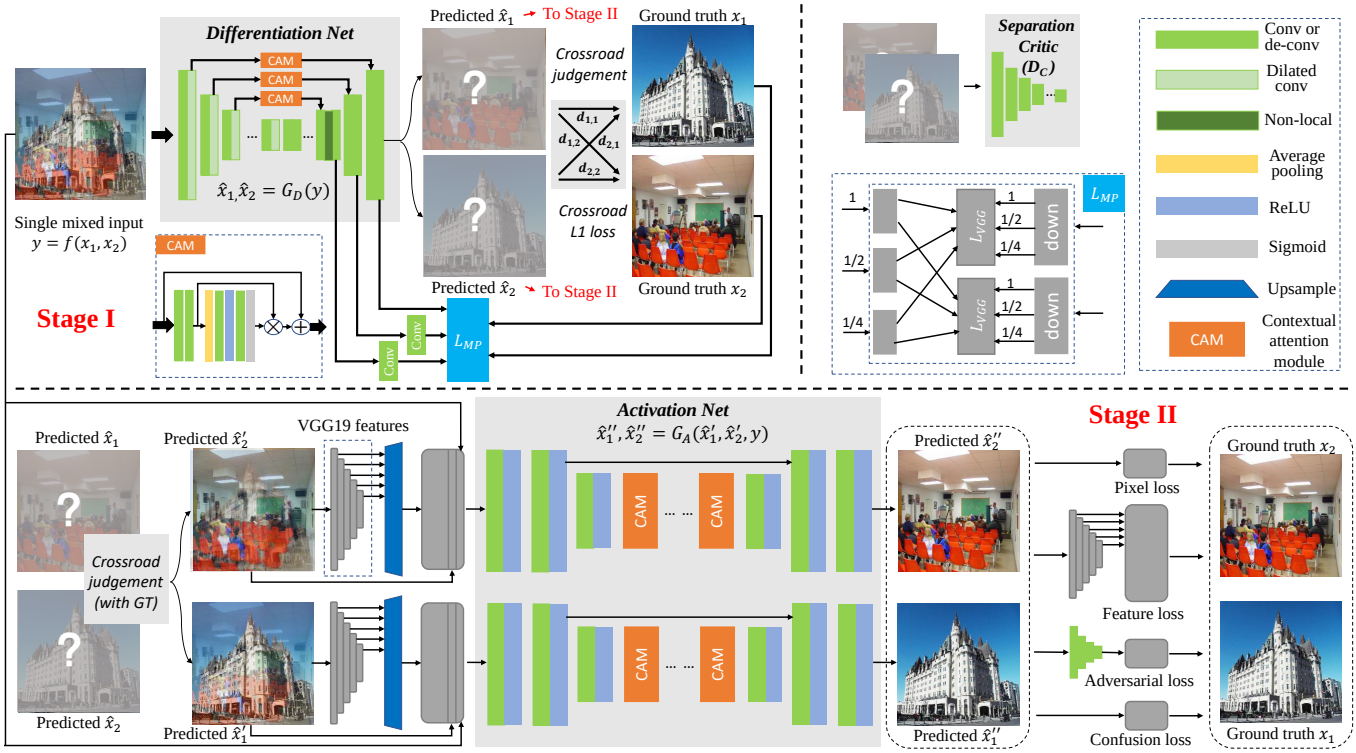


Fig. 1. An overview of the proposed method. The proposed method consists of two stages, including "Stage I: development stage" and "Stage II: rivalry stage". In the development stage, the differentiation net G_D tries to distinguish the mixed input as much as possible, which is trained under three losses including "crossroad $\mathcal{L}_1$", "separation critic D_C", and "multi-scale crossroad perceptual loss L_{MP}", respectively. In the rivalry stage, the activation net G_A tries to activate one layer from the mixed input by combining the information from "Stage I" and reduce the confusion, which is trained under four losses including "pixel loss", "feature loss", "adversarial loss", and "confusion loss", respectively.

removal mainly focused on using multi-image approaches [14] and hand-crafted priors such as smoothness prior [35], ghosting cues [36], gradient sparsity constraint [37], *etc.*, to reduce the reflection effect. Recently, deep learning-based methods have been used for the single image reflection removal task [38] and have achieved great improvement. Fan *et al.* [15] have proposed a two stage Convolutional Neural Network (CNN), which first predicts the edge information of the transmission layer image and then utilizes this information to generate the reflection-free image. Yang *et al.* [39] have designed a bidirectional network which alternately estimates the transmission layer and the reflection layer to remove the reflection. Zhang *et al.* [40] have employed the perceptual loss [41] to improve the quality of restored images. Wei *et al.* [4] have proposed an alignment invariant loss to resolve the misaligned problem in training with real world data. Li *et al.* [42] have designed a two-stage reflection removal framework via reflection-aware guidance. Compared to this two-stage model, our model focuses on a more basic task and can well handle many related tasks.

2) Single image rain removal. A rainy image can be viewed as the superimposition of a clean background layer and a rain streak layer [43]. Some early works have considered rain streaks as a kind of high frequency noise and used low-rank constraints [44] or sparse coding [45] to solve the problem. Recently, deep learning-based approaches have emerged for single image rain removal [46]–[48] and achieved impressive restoration performance. Fu *et al.* [46] have introduced a three layer CNN on the high frequency domain of the image to remove rain streaks. Yang *et al.* [49] have proposed a continuous process to sequentially detect, estimate, and remove rain streak. Zhang *et al.* [50] have applied generative adversarial network (GAN) mechanism and perceptual loss function into the single image rain removal task. Jiang *et al.* [10] have proposed a multi-scale progressive fusion network (MSPFN) to exploit the correlated information of rain streaks across scales for single image deraining.

3) Single image shadow removal. An image with shadow can also be viewed as the superimposition of a clean image and a shadow mask. Conventional shadow removal methods removed shadows based on hand-crafted features, such as illumination invariant assumptions [51], edge and pixel features [52], region level cues [53], *etc.* Recently, some deep learning based methods have made significant performance improvement on the shadow removal task. Qu *et al.* [17] have proposed a multi-context architecture named DeshadowNet for single image shadow removal. Wang *et al.* [8] have designed a stacked conditional Generative Adversarial Network (ST-CGAN) for joint learning shadow detection and shadow removal. Ding *et al.* [7] have proposed an attentive recurrent generative adversarial network (ARGAN) to progressively remove shadow.

4) Illumination correction. A low-light image can be seen as the superimposition of a normal-light image with some perturbations (such as noise, exposure, *etc.*) [54]. Traditional low-light enhancement methods have adjusted the illumination uniformly based on histogram equalization [55] or adjusted adaptively based on retinex methods [56], which may result

in under-exposed or over-exposed local details. Recently, some deep learning based methods have achieved better performance for low-light enhancement [57], such as GAN based method [58] or two-stage semi-supervised method [59], [60].

5) Unified framework. The works mentioned above only focused on one specific application, and two recent studies [20], [61] have made attempts to solve different tasks of image restoration [62]–[64] in a unified framework. Zou *et al.* [20] have proposed a superimposed image decomposition method based on adversarial supervision, which can be generalized to various image separation tasks. Feng *et al.* [61] have designed a unified framework to recover the background information of images with various noise based on deep noise estimation.

III. Proposed Method

In this section, we describe the proposed framework in detail. In the first stage, we aim at simulating the development stage of human vision when viewing a single superimposed image, which tries to distinguish the two layers of the superimposed image. For the second stage, we aim at simulating the monocular rivalry stage of human vision, which tries to activate one layer and suppress another layer. Fig. 1 demonstrates the overview of the proposed framework.

A. Stage I: Development Stage

We first introduce the development stage. Suppose x_1 and x_2 represent two individual images, and $y = f(x_1, x_2)$ denotes the mixture of them, where $f(\cdot)$ could be a linear or non-linear function. Our objective is to distinguish $\hat{x}_1$ and $\hat{x}_2$ from a single mixed input y as follows:

$$\hat{x}_1, \hat{x}_2 = G_D(y), \tag{1}$$

where G_D denotes the proposed differentiation net (DiNet).

1) Network Architecture: The architecture of the proposed differentiation network is illustrated in the "Stage I" part in Fig. 1. The DiNet G_D is built based on the configuration of the "U-Net" [20], [65]. Intuitively, this autoencoder with skip connection structure can better extract and disentangle the superimposed features and coarsely recover two layers. Since it may introduce unexpected artifacts and arise stability issues [66], we do not use the batch-normalization in G_D. As mentioned above, it is important to consider enlarging the receptive field of the network to distinguish visual confusion. Therefore, for the first four convolutional layers in the encoder, we enlarge the receptive field by adding a dilated convolutional layer after each convolutional layer. Moreover, we use a non-local layer [67] in the decoder part to better perceive the whole image.

Inspired by the attention mechanism under development stage of human vision, we also consider the context between channels (image features) for the differentiation net. In this paper, we leverage the contextual attention module [25] (illustrated as CAM in Fig. 1, *a.k.a*, channel attention module) to introduce global contextual information across channels for better disentangling superimposed features. Let $U \in \mathbb{R}^{H\times W\times C}$ denote the input feature block of the CAM, where $H \times W$ indicates the spatial scale, C represents the channel number, we first refine the feature block U with two convolutional layers to produce $V \in \mathbb{R}^{H\times W\times C}$. Then we apply a global average pooling $\mathcal{A}$ to each feature channel of V to obtain the channel-specific descriptor vector $Z \in \mathbb{R}^{1\times 1\times C}$, which can be expressed as $Z = \mathcal{A}(V)$. After passing this channel descriptor vector Z through an attention module, which includes a down-sampling linear convolution layer, a ReLu layer, an up-sampling linear convolution layer and a sigmoid layer, the attention descriptor $S \in \mathbb{R}^{1\times 1\times C}$ for the feature block V is generated. This attention descriptor S serves as the channel-specific gate for calibrating the feature block V via: $\hat{V} = S \cdot V$. Finally, a residual architecture with reference to the input for easier optimization is implemented by: $\hat{U} = \hat{V} + U$, to produce the output feature vector $\hat{U}$ of the CAM.

2) Objective Function: The objective function of DiNet contains three terms: a crossroad $\mathcal{L}_1$ loss, a separation critic, and a multi-scale perceptual loss.

Crossroad $\mathcal{L}_1$ Loss. The DiNet G_D is trained to minimize the distance between its outputs $(\hat{x}_1, \hat{x}_2)$ and their ground truths (x_1, x_2). Since the order of the decomposition outputs is not specified, we use the crossroad $\mathcal{L}_1$ loss [20] to measure the pixel-wise distance between the predicted outputs and the ground truths, which is defined as:

$$\begin{aligned} &l_{\text{cross}}((\hat{x}_1, \hat{x}_2), (x_1, x_2)) \\ &= \min\{d_{1,1} + d_{2,2}, d_{1,2} + d_{2,1}\}, \end{aligned} \tag{2}$$

where $d_{i,j} = \|\hat{x}_i - x_j\|, i, j \in \{1, 2\}$. Hence, this objective function on an entire dataset can be expressed as:

$$\mathcal{L}_{\text{cross}} = \mathbb{E}_{x_i\sim p_i(x_i)}\{l_{\text{cross}}((\hat{x}_1, \hat{x}_2), (x_1, x_2))\}, \tag{3}$$

in which $i \in \{1, 2\}$, and $p_i(x_i)$ indicates the distribution of the image data.

Separation Critic. To further improve the separation performance, a decomposition prior learned through an adversarial training is introduced [20], which tries to distinguish the outputs $(\hat{x}_1, \hat{x}_2)$ and a pair of clean images (x_1, x_2). The discriminator D_C is defined as:

$$\begin{aligned} \mathcal{L}_{\text{critic}}^{D_C} &= \mathbb{E}_{x_i\sim p_i(x_i)}\{\log D_C(x_1, x_2)\} \\ &+\mathbb{E}_{\hat{x}_i\sim p_i(\hat{x}_i)}\{\log(1 - D_C(\hat{x}_1, \hat{x}_2))\}, \end{aligned} \tag{4}$$

where $D_C(x, y)$ is the probability that the pair (x, y) is a well-separated (clean) image pair. The structure of the discriminator D_C is the same as pix2pix [68], of which the two input images are simply concatenated at the input end. The loss function of the generator G_D is defined as:

$$\mathcal{L}_{\text{critic}}^{G_D} = \mathbb{E}_{\hat{x}_i\sim p_i(\hat{x}_i)}\{-\log(D_C(\hat{x}_1, \hat{x}_2))\}. \tag{5}$$

The adversarial training of G_D and D_C is a minimax optimization process, where G_D tries to minimize the objective function while D_C tries to maximize it.

Multi-scale Crossroad Perceptual Loss. Multi-scale losses are proved to be effective in optimizing image decomposition tasks such as de-raining [26] and reflection removal [18]. A multi-scale loss first extracts features from different decoder layers and then feeds them into a convolutional layer to form outputs at different scales. By comparing the perceptual (feature) distance between these multi-scale outputs to those

real images with the corresponding scales, multi-scale perceptual losses can be obtained, which enable the net to capture more contextual information from various scales. We adopt the perceptual distance over different scales rather than other loss functions in order to utilize both low-level and high-level information. However, the order of the decomposition outputs of multiple scales is also not specified. Therefore, we introduce multi-scale *crossraod* perceptual losses in this paper. We first propose a crossroad judgement to match the predicted outputs to the ground truths:

$$\begin{aligned} &\hat{x}'_1 = \hat{x}_i, \hat{x}'_2 = \hat{x}_j, \\ &\textbf{s.t.}\ \min\{d_{i,1} + d_{j,2}\}, \end{aligned} \tag{6}$$

where $d_{i,1} = \|\hat{x}_i - x_1\|$, $d_{j,2} = \|\hat{x}_j - x_2\|$, $i, j \in \{1, 2\}$, $i \neq j$, $\hat{x}_i, \hat{x}_j$ are the predicted outputs of the DiNet, x_1, x_2 are the ground truths, and $\hat{x}'_1, \hat{x}'_2$ are the outputs after the crossroad judgement. By crossly judging the distance between the outputs of the DiNet and the ground truths, we can match the pair $(\hat{x}_i, \hat{x}_j)$ to the ground truth pair (x_1, x_2), and then match the pair $(\hat{x}'_1, \hat{x}'_2)$ and (x_1, x_2) in order. Then we define the multi-scale crossroad perceptual loss as:

$$\mathcal{L}_{\text{MP}} = \sum_{k=1}^{M} (\lambda_k (\mathcal{L}_{\text{VGG}}(\hat{x}'_{1_k}, x_{1_k}) + \mathcal{L}_{\text{VGG}}(\hat{x}'_{2_k}, x_{2_k}))), \tag{7}$$

where $\hat{x}'_{1_k}, \hat{x}'_{2_k}$ indicate the k-th outputs extracted from the decoder layers, x_{1_k}, x_{2_k} indicate the ground truths which have the same scale as $\hat{x}'_{1_k}$ and $\hat{x}'_{2_k}$, and λ_k indicate the constraints for different scales. $\mathcal{L}_{\text{VGG}}$ (perceptual (feature) loss [41]) is defined as $\mathcal{L}_{\text{VGG}}(x, y) = \sum_l \omega_l \|\phi_l(x) - \phi_l(y)\|_1$, where ϕ_l indicates the l-th layer in the VGG network, and $\{\omega_l\}$ are used to balance the terms in the loss function. Specifically, in our implementation, we set $M = 3$, and the outputs of the last 1^{st}, 3^{rd}, 5^{th} layers are used, whose sizes are 1, $\frac{1}{2}$, $\frac{1}{4}$ of the original size, respectively. λ_1, λ_2, λ_3 are set to 1, 0.8, 0.6, respectively. We use VGG-19 [69] as the backbone of the perceptual loss function, where the weights ω_l are the same as that in [40].

Our overall loss function of DiNet is:

$$\mathcal{L}_{\text{di}} = \alpha_1 \mathcal{L}_{\text{cross}} + \alpha_2 \mathcal{L}_{\text{critic}} + \alpha_3 \mathcal{L}_{\text{MP}}, \tag{8}$$

where α_1, α_2, α_3 control the balance among different components of the loss function, which are empirically set to 1, 0.0001, and 0.1, respectively.

B. Stage II: Rivalry Stage

We then simulate the monocular rivalry stage of human vision. For a mixed input y, during one period of monocular rivalry, only one layer is activated. To this end, we first pass $(\hat{x}_1, \hat{x}_2)$ through a crossroad judgement module as described in Section III-A2 and get the pair $(\hat{x}'_1, \hat{x}'_2)$ matched in order with the ground truth (x_1, x_2) to decide which layer to activate. Then we feed $(\hat{x}'_1, \hat{x}'_2)$ with the mixed input y together to the activation net to activate one layer and suppress another layer of the mixed input:

$$\hat{x}''_1, \hat{x}''_2 = G_A(\hat{x}'_1, \hat{x}'_2, y), \tag{9}$$

where G_A is the proposed activation net (AcNet).

1) Network Architecture: The architecture of the proposed activation network (AcNet) is built based on the Resnet generator [41] as illustrated in the "Stage II" part in Fig. 1. A dual pathway parallel net is designed, of which two pathways share weights with each other. For the obtained $\hat{x}'_1$ or $\hat{x}'_2$, we first extract the hypercolumn features [70] from a pretrained VGG-19 network [69], and then concatenate these features with $\hat{x}'_1$ or $\hat{x}'_2$ as an augmented network input. This augmentation strategy for the input enables the network to learn more semantic clues from the image [40]. We further concatenate the single mixed input y to the input which aims at utilizing the complete texture information from y. Then we feed the input into our AcNet. The AcNet contains 7 cascaded CAM blocks of which the architecture is the same with that in "Stage I". Moreover, we add the skip connection between the second convolutional layer and the second to last convolutional layer as the residual structure for easier optimization. Through the AcNet, we can reduce the confusion and improve the quality of the outputs from the DiNet.

2) Objective Function: The objective function of the AcNet contains four terms: a pixel loss, a feature loss, an adversarial loss, and a confusion loss.

Pixel Loss. To ensure that the outputs are as close to the ground truths as possible, we utilize $\mathcal{L}_1$ loss to measure the pixel-wise distance between them, which is defined as:

$$\mathcal{L}_{\text{pixel}} = \mathbb{E}_{(\hat{x}''_i, x_i) \sim p_i(\hat{x}''_i, x_i)} \{\mathcal{L}_1(\hat{x}''_i, x_i)\}. \tag{10}$$

Feature Loss. We compute the feature loss by feeding the predicted output and the ground truth through a pretrained VGG-19 network respectively, then compute the $\mathcal{L}_1$ distance between the selected feature layers. The feature loss in this work is defined as:

$$\mathcal{L}_{\text{feat}} = \mathbb{E}_{(\hat{x}''_i, x_i) \sim p_i(\hat{x}''_i, x_i)} \{\mathcal{L}_{\text{VGG}}(\hat{x}''_i, x_i)\}, \tag{11}$$

where $\mathcal{L}_{\text{VGG}}$ is the same as that mentioned in Section III-A2.

Adversarial Loss. To encourage the predicted output to be as realistic as the ground-truth image layer, an adversarial loss [68] is used to improve the realism of the predicted output. The loss function of the discriminator D is defined as:

$$\begin{aligned} \mathcal{L}^D_{\text{adv}} = &\mathbb{E}_{(\hat{x}''_i, y_i) \sim p_i(\hat{x}''_i, y_i)} \{\log D(\hat{x}''_i, y_i)\} \\ &- \mathbb{E}_{(x_i, y_i) \sim p_i(x_i, y_i)} \{\log D(x_i, y_i)\}, \end{aligned} \tag{12}$$

and the loss function of the generator G is defined as:

$$\mathcal{L}^G_{\text{adv}} = -\mathbb{E}_{(\hat{x}''_i, y_i) \sim p_i(\hat{x}''_i, y_i)} \{\log D(\hat{x}''_i, y_i)\}. \tag{13}$$

The structure of discriminator is also the same as pix2pix [68]. It is worth nothing that the function of this adversarial loss is to improve the quality of the output, while the Eq. (4) and Eq. (5) is used for better separating two layers.

Confusion Loss. We further propose a confusion loss to reduce the confusion content from the non-target image which is calculated in gradient domain. As discussed in [40], the edges of two ground truths are unlikely to overlap, however, numerically, the correlation between the gradient maps of these two images is still countable. In this paper, we modify the exclusion loss proposed in [40] and calculate the residual confusion loss. We first calculate the correlation between the output and the non-target ground truth in gradient domain,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3172882, IEEE Transactions on Multimedia

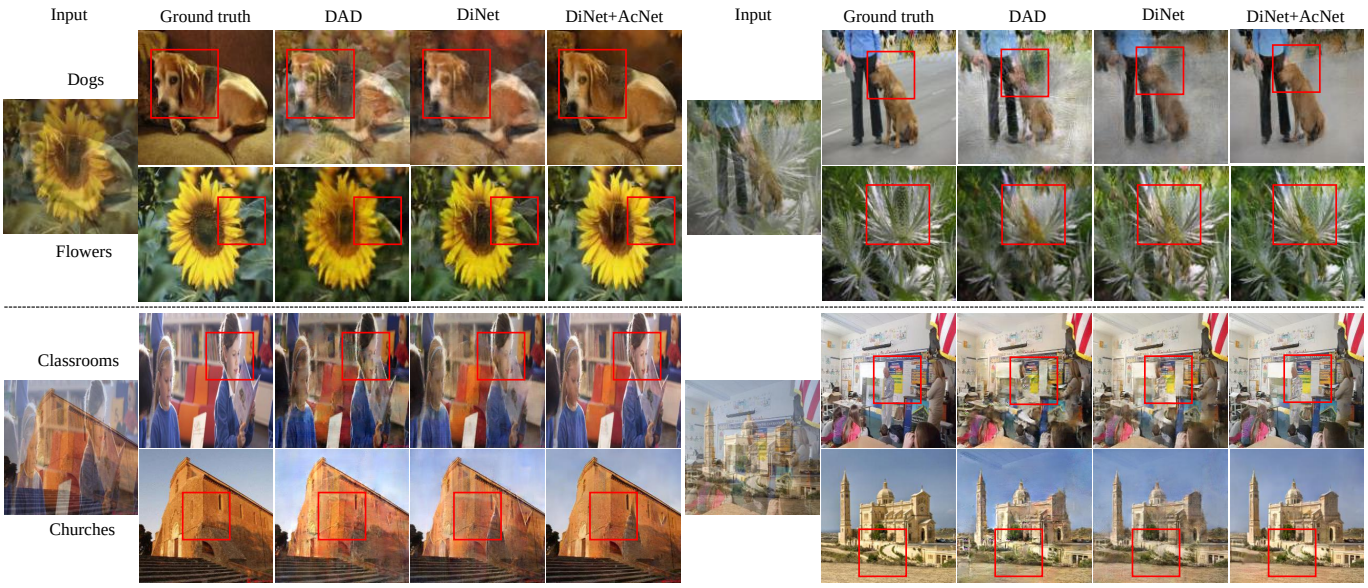


Fig. 2. Qualitative comparison of superimposed image decomposition on two mixing datasets. Above dashed line: superimposed image decomposition on Dogs+Flwrs. Under dashed line: superimposed image decomposition on LSUN mixed dataset. Our methods can separate the two layers of a superimposed image with less artifacts from the undesired layer, particularly in the regions indicated by bounding boxes. Best viewed in zoom-in mode.

TABLE I
PERFORMANCE (PSNR / SSIM) OF DIFFERENT METHODS FOR SUPERIMPOSED IMAGE SEPARATION ON TWO MIXING DATASETS: 1) DOGS [71] + FLWRS [72], AND 2) LSUN CLASSROOM + LSUN CHURCH [73]. $*$: TRAINED ON IMAGENET. $\#$: TRAINED ON DEFAULT DATASETS. (WE **BOLD** THE BEST RESULTS AND UNDERLINE THE SECOND-BEST RESULTS. THE SAME HIGHLIGHT METHOD IS USED IN THE FOLLOWING TABLES.)

Methods	Dogs+Flwrs	LSUN
Levin *et al.* [11] (TPAMI'07)	10.54 / 0.444	10.46 / 0.366
Double-DIP [19] (CVPR'19)	14.70 / 0.661	13.83 / 0.590
DAD [20] (CVPR'20) $*$	23.32 / 0.803	21.63 / 0.773
DAD [20] (CVPR'20) $\#$	25.51 / 0.849	26.32 / 0.883
DiNet	26.65 / 0.876	27.13 / 0.901
DiNet + AcNet	**28.93 / 0.921**	**30.47 / 0.947**

TABLE II
ABLATION STUDIES FOR THE ARCHITECTURE AND LOSSES OF DINET ON TWO MIXING DATASETS: 1) DOGS [71] + FLWRS [72], AND 2) LSUN CLASSROOM + LSUN CHURCH [73]. THE PERFORMANCE IS FORMATTED AS PSNR / SSIM

Methods	Dogs+Flwrs	LSUN
basenet	25.55 / 0.850	26.05 / 0.880
w/o DC	26.54 / 0.872	27.20 / 0.902
w/o CA	26.44 / 0.868	26.86 / 0.896
w/o SA	26.48 / 0.870	27.13 / 0.901
w/o $\mathcal{L}_{\text{critic}}$	26.18 / 0.864	26.96 / 0.894
w/o $\mathcal{L}_{\text{MP}}$	26.17 / 0.864	26.93 / 0.892
rp $\mathcal{L}_{\text{MP}}$ with $\mathcal{L}_{\text{P}}$	26.26 / 0.867	27.01 / 0.894
all combined	**26.65 / 0.876**	**27.23 / 0.902**

and the correlation between the target ground truth and the non-target ground truth in gradient domain, respectively. Then we formulate the confusion loss as the residual correlation by calculating the difference between these two correlations:

$$\mathcal{L}_{\text{conf}} = \mathbb{E}_{(\hat{x}_i'', x_i)\sim p_i(\hat{x}_i'', x_i)} \{\|\psi(\hat{x}_1'', x_2) - \psi(x_1, x_2)\|_2 + \|\psi(\hat{x}_2'', x_1) - \psi(x_2, x_1)\|_2\}, \tag{14}$$

$$\psi(x, y) = \sigma(\lambda_x |\nabla x|) \odot \sigma(|\nabla y|), \tag{15}$$

where ∇ is the gradient, $\lambda_x = \dfrac{\sum |\nabla y|}{\sum |\nabla x|}$ is used as the normalization factor, σ denotes the sigmoid function, $\odot$ indicates the element-wise multiplication. By minimizing the confusion, we aim at reducing the confusion content from the non-target ground truth while keeping the texture information from the target ground truth.

Our overall loss function for AcNet is:

$$\mathcal{L}_{\text{ac}} = \beta_1 \mathcal{L}_{\text{pixel}} + \beta_2 \mathcal{L}_{\text{feat}} + \beta_3 \mathcal{L}_{\text{adv}} + \beta_4 \mathcal{L}_{\text{conf}}, \tag{16}$$

where weighting coefficients β_1, β_2, β_3, β_4 are empirically set to 1, 0.1, 0.0001, 0.1, respectively.

C. Implementation Details

We implement the proposed framework in Pytorch on a server with an Nvidia Geforce RTX 2080 Ti GPU. Generally, the DiNet is trained for 100 epochs first with a batch size of 2, using the Adam optimizer [74]. Then, we freeze the weights of the DiNet, and only train the AcNet for 100 epochs with the batch size of 2, using the Adam optimizer, too. The learning rates of two networks are both set to 0.0001. For some datasets, the DiNet and the AcNet are trained for more than 100 epochs, respectively. The learning rate is decayed by ten times from 100 epochs. We set α_2 and β_3 to 0 for the first 5 epochs, and then set them to 0.0001 for the rest epochs, respectively.

IV. EXPERIMENTAL VALIDATION

We evaluate the proposed method on 5 tasks, including 1) superimposed image separation, 2) single image reflection removal, 3) single image rain removal, 4) single image shadow removal, and 5) single image low-light enhancement. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3172882, IEEE Transactions on Multimedia

TABLE III
ABLATION STUDIES FOR THE ARCHITECTURE AND LOSSES OF ACNET ON THE DOGS [71] + FLWRS [72] DATASET.

Methods	UNet 1	UNet 2	w/o MI	1CAM	3CAM	4CAM	5CAM	6CAM	w/o $\mathcal{L}_{adv}$	w/o $\mathcal{L}_{feat}$	w/o $\mathcal{L}_{conf}$	all combined
PSNR	27.24	27.28	27.34	27.27	27.63	28.34	28.61	28.93	28.92	27.92	28.82	**29.03**
SSIM	0.892	0.892	0.890	0.895	0.901	0.913	0.916	0.922	0.920	0.903	0.918	**0.922**

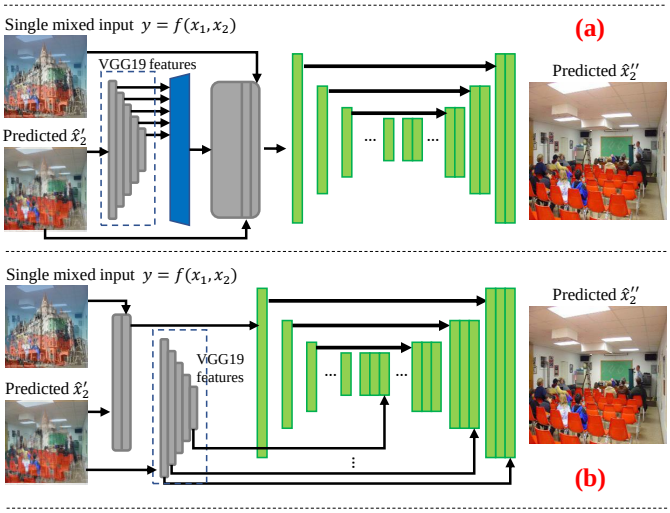


Fig. 3. The architecture of two UNets in Table III.

experimental settings and results are described and analyzed in detail as follows.

A. Separating Superimposed Images

Experimental settings. As the basic task of this paper, we first evaluate the performance of the proposed method on the task of separating superimposed images. We follow the experimental protocol in [20] and evaluate the proposed method on two datasets of mixed image decomposition: 1) Stanford-Dogs (Dogs) [71] + VGG-Flowers (Flwrs) [72], and 2) LSUN Classroom + LSUN Church [73]. Following Zou *et al.* [20], we use the original training/testing split in these datasets [71]–[73] to conduct the experiments. During training, we randomly choose two images (x_1, x_2) from the two subsets of one group then linearly mix them with a value α to get the mixed input y as: $y = f(x_1, x_2) = \alpha x_1 + (1-\alpha)x_2$, where $\alpha \in [0.4, 0.6]$. During evaluation, we use the same sequence as in [20] to mix two images with a constant α value of 0.5. Overall, we randomly produce 6149 pairs of Dogs+Flwrs and 126227 pairs of LSUN Classroom+Church for each epoch of the training process, and fixedly generate 1020 pairs of Dogs+Flwrs and 300 pairs of LSUN Classroom+Church for the testing process. The DiNet and AcNet are trained for 100 epochs on the Dogs+Flwrs dataset, and trained for 20 epochs on the LSUN Classroom+Church dataset (due to the large number of images in this dataset), respectively.

We compare the proposed model with three popular methods for single superimposed image decomposition, including one user-interaction needed method (Levin *et al.*) [11], one unsupervised method (Double-DIP) [19], and one supervised method (DAD) [20].

Results. Table I presents the quantitative comparison results of different methods for superimposed image decomposition on two mixing datasets in terms of PSNR and SSIM. It manifests that our model achieves the best performance in terms of both metrics on two datasets. We further observe that the performances of *Levin et al.* and *Double-DIP* are inferior to other methods, which is ascribed to that the former one heavily needs user-interaction and the latter one is an unsupervised method, and both of them can hardly work on separating the superimposition of two complex images. It is worth mentioning that our DiNet achieves significant improvement compared with *DAD*, and the DiNet+AcNet distinctly outperforms these one-stage methods. To gain more insight into the performance comparisons, we show some visualization examples of the separation results in Fig. 2. The distinguished regions are highlighted with red rectangles. It qualitatively manifests that our DiNet separates the superimposed image better compared to *DAD* with less artifacts. Moreover, after the process of AcNet, the residual information from the non-target image can be reduced and the overall image color can be corrected, which further justifies the superiority of our two-stage method.

B. Ablation Studies

We further conduct ablation studies to investigate the effect of each component in our DiNet and AcNet, respectively.

Ablation studies for DiNet. We first perform ablation experiments on seven variants of the DiNet, which includes: 1) *basenet*, whose structure is similar to the UNet in [68], [76], and loss functions are $\mathcal{L}_{\text{cross}}$ and $\mathcal{L}_{\text{critic}}$, 2) *w/o DC*, which means without dilated convolutional layer, 3) *w/o CA*, which indicates without channel attention module, 4) *w/o SA*, which represents without spatial attention module, 5) *w/o* $\mathcal{L}_{\text{critic}}$, which means without adversarial loss $\mathcal{L}_{\text{critic}}$ 6) *w/o* $\mathcal{L}_{\text{MP}}$, which implies without the multi-scale crossroad perceptual loss $\mathcal{L}_{\text{MP}}$, and 7) *rp* $\mathcal{L}_{\text{MP}}$ *with* $\mathcal{L}_{\text{P}}$, which denotes replacing the multi-scale crossroad perceptual loss $\mathcal{L}_{\text{MP}}$ with only one crossroad perceptual loss $\mathcal{L}_{\text{P}}$.

Table II shows the results of the ablation experiments for DiNet. To investigate the contribution of each component in DiNet, we first compare the performances of *w/o DC*, *w/o CA*, and *w/o SA*. It can be observed that all these three modules have benefits to the final performance, and the channel attention module contributes more than other two modules. Intuitively, the channel attention module promotes the learning of feature attention, thus better help disentangle the features of the superimposed image. Table II also presents the performance of DiNet without the adversarial loss $\mathcal{L}_{\text{critic}}$ and the multi-scale crossroad perceptual loss $\mathcal{L}_{\text{MP}}$. We observe that both losses yield notable performance improvement while

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3172882, IEEE Transactions on Multimedia

(a) Input (b) Ground truth (c) Zhang *et al.* (d) BDN (e) DAD (f) DiNet (Ours) (g) DiNet+AcNet (Ours)

Fig. 4. Qualitative comparison of different methods for reflection removal on two test datasets. 1st, 2nd, and 3rd rows: on BDN [39] dataset. 4th, 5th, and 6th rows: on the dataset of Zhang *et al.* [40]. Our methods are able to restore clearer background images with less artifacts than other methods.

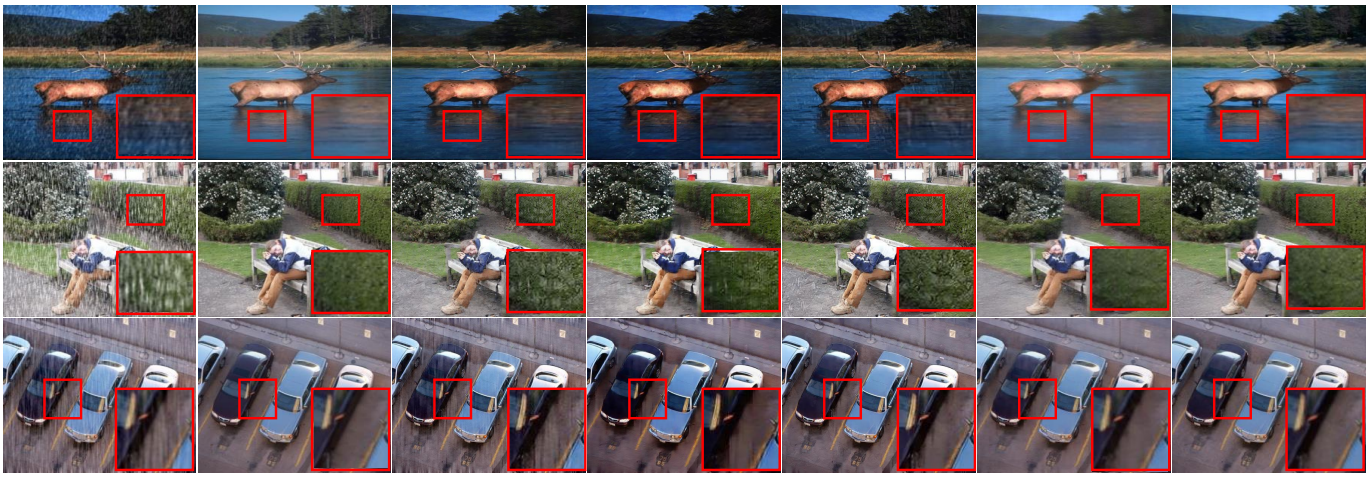
TABLE IV
QUANTITATIVE EVALUATION (PSNR / SSIM) OF DIFFERENT METHODS FOR IMAGE REFLECTION REMOVAL ON TWO CHALLENGING SYNTHETIC DATASETS [75] AND [39].

Methods	Dataset [75] Focused set	Dataset [75] Defocused set	Dataset [75] Ghosting set	Methods	Dataset [39]
CEILNet [15] (ICCV'17)	19.52 / 0.742	20.12 / 0.735	19.68 / 0.753	Li & Brown [35] (CVPR'14)	16.46 / 0.745
Zhang *et al.* [40] (CVPR'18)	17.09 / 0.712	18.10 / 0.758	17.88 / 0.738	SIRP [37] (CVPR'17)	19.18 / 0.760
BDN [39] (ECCV'18)	14.25 / 0.632	14.05 / 0.639	14.78 / 0.660	CEILNet [15] (ICCV'17)	19.80 / 0.782
RmNet [75] (CVPR'19)	21.06 / 0.770	22.89 / 0.840	21.00 / 0.780	BDN [39] (ECCV'18)	23.11 / 0.835
DAD [20] (CVPR'20)	22.80 / 0.871	23.19 / 0.891	23.26 / 0.881	DAD [20] (CVPR'20)	23.18 / 0.877
DiNet (proposed)	24.19 / 0.893	24.78 / 0.916	24.80 / 0.904	DiNet (proposed)	24.21 / 0.899
DiNet + AcNet (proposed)	**24.94 / 0.908**	**26.07 / 0.938**	**25.70 / 0.919**	DiNet + AcNet (proposed)	**25.56 / 0.916**

$\mathcal{L}_{\text{MP}}$ has relatively larger contribution. To further verify our supposition that learning decomposition earlier will contribute to the final performance, we replace the multi-scale crossroad perceptual loss $\mathcal{L}_{\text{MP}}$ with the crossroad perceptual loss $\mathcal{L}_{\text{P}}$ which only acts on the last layer. Comparing the performance of *rp* $\mathcal{L}_{\text{MP}}$ *with* $\mathcal{L}_{\text{P}}$ and the performance of *w/o* $\mathcal{L}_{\text{MP}}$, the improvement caused by $\mathcal{L}_{\text{P}}$ is not obvious. Intuitively, the $\mathcal{L}_{\text{MP}}$ may suppress the congestion from the non-target image.

Ablation studies for AcNet. We then perform ablation experiments on eleven variants of the AcNet and present the results in Table III. To verify the effectiveness of the architecture, we first compare our AcNet with two UNet structures as shown in Fig. 3. We notice that adopting UNet as the second stage cannot yield notable improvement compared to our AcNet. We then analyze the architecture of our AcNet. We observe significant improvement by incorporating the mixed input to activate one layer rather than only refining the output from the last stage, as indicated by *w/o MI*. In this paper, 7 CAMs are used in the AcNet, so we also consider reducing the number of CAMs to investigate the performance of AcNet. As shown by the performances of *1CAM*, *3CAM*, *4CAM*, *5CAM*, *6CAM*, and *all combined* in Table III, we notice that when

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3172882, IEEE Transactions on Multimedia



(a) Input (b) Ground truth (c) RESCAN (d) DIDMDN (e) PreNet (f) DiNet (Ours) (g) DiNet+AcNet (Ours)

Fig. 5. Qualitative comparison of different methods for single image rain removal on Test100 [50] dataset.

TABLE V
QUANTITATIVE RESULTS (PSNR / SSIM) OF DIFFERENT METHODS FOR REFLECTION REMOVAL ON A REAL DATASET [40].

Methods	Real20 [40]	Wild [3]
CEILNet [15] (ICCV'17)	19.04 / 0.762	22.14 / 0.819
Zhang *et al.* [40] (CVPR'18)	21.30 / 0.821	21.52 / 0.829
BDN [39] (ECCV'18)	20.06 / 0.738	22.34 / 0.821
ERRNet [4] (CVPR'19)	22.80 / 0.803	24.16 / 0.847
DAD [20] (CVPR'20)	22.36 / 0.846	24.80 / 0.922
DMGN [61] (TIP'21)	23.05 / 0.823	25.18 / 0.894
Li *et al.* [42] (arxiv'21)	22.95 / 0.793	25.52 / 0.880
DiNet (proposed)	23.11 / 0.870	25.56 / 0.926
DiNet + AcNet (proposed)	**23.80 / 0.877**	**25.69 / 0.929**

the number of CAM modules is larger than 5, increasing the number of CAMs has less effect on the performance improvement. Thus 7 CAMs are adopted in this paper. We also analyze the contribution of each loss function in the AcNet, as indicated by *w/o* $\mathcal{L}_{\text{adv}}$, *w/o* $\mathcal{L}_{\text{feat}}$, and *w/o* $\mathcal{L}_{\text{conf}}$ in Table III. We notice that all loss functions have contributions to promote the performance, while the perceptual loss $\mathcal{L}_{\text{feat}}$ contributes the most to the improvement.

Model size comparison with [20]. The model size of our DiNet in the first stage is about 61.04 M. Compared with Zou *et al.* [1], of which the model size is about 54.1 M, our model can achieve much better results. Moreover, the AcNet in the second stage is about 11.58 M, which is a more slight but efficient module.

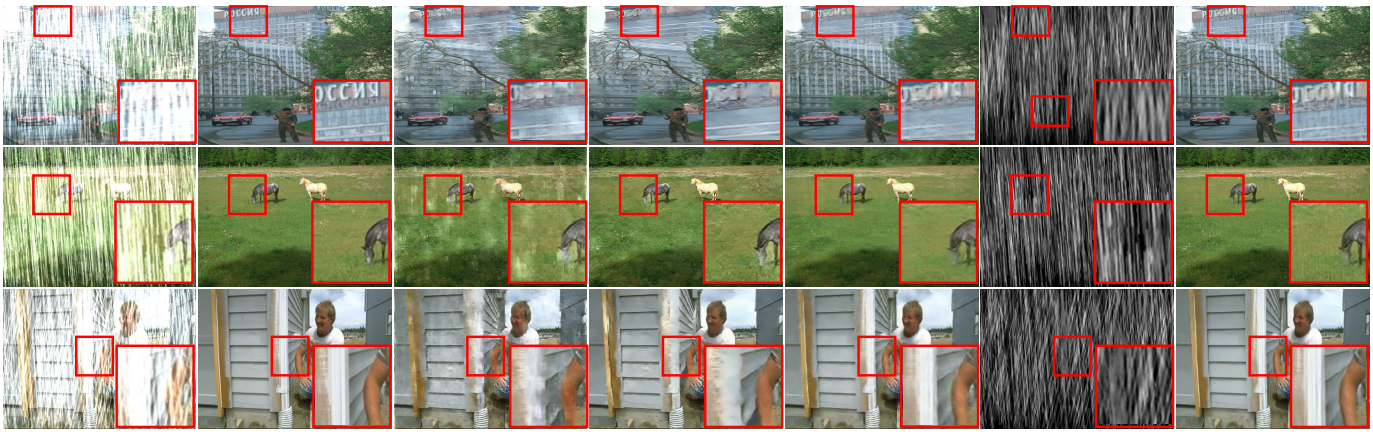
C. Application: Single Image Reflection Removal

Experimental settings. We conduct single image reflection removal experiments on two synthetic datasets [39], [75] and two real datasets [3], [40]. The synthetic dataset *Syn* [75] contains three types of reflections including "focused", "defocused" and "ghosting", which are synthesized using adversarial training. Each reflection type includes 4000 images for training and 100 images for testing, which results in 12000 training images and 300 test images in total. The dataset *BDN* [39] synthesized the images with reflections by linearly mixing the clear transmission layer and blurred reflection layer, which contains 50000 training images and 400 test images. To validate the generalization ability of our model to real cases, two real datasets including *real20* [40] and *wild* [3] are involved in the experiments. We follow the common training/test methods which have been widely used in the literature [4], [20], [40], [61] to conduct the experiments. The training data consists of two parts, which include synthetic image pairs randomly synthesized from clean Flickr images and real image pairs randomly cropped from real-world images with reflections. The two test datasets real20 [40] and wild [3] contain 20 real-world image pairs with reflections across various scenes and 55 image pairs collected from the wild scenes, respectively. Since the ground truths of blurred reflection images are usually unavailable, we simply set the ground truth of the second output as a "zero image" [20] and only train one pathway of the AcNet during training.

We compare our model with state-of-the-art methods for single image reflection removal including: *Li & Brown* [35], which removes reflections using relative smoothness; *SIRP* [37], that suppresses reflections based on a Laplacian data fidelity term and a sparsity term; *CEILNet* [15], which presents a cascaded edge and image learning network for reflection removal; *Zhang et al.* [40], which proposes to solve reflection removeal problem by perceptual loss; *BDN* [39], which improves the restoration quality by a bidirectional network that alternately estimates the transmission image and the reflection image; *RmNet* [75], which uses GAN to remove reflection; *ERRNet* [4], which focuses on resolving the misaligned problem in real training data and leveraging the multi-scale context; *DAD* [20], which proposes a unified framework for separating superimposed images by GAN; and *DMGN* [61], which designs a unified framework for superimposed image restoration by estimating the noise mask.

Results. Table IV presents the quantitative results of different models for single image reflection removal on two synthetic datasets [39], [75]. It manifests that both of our DiNet and DiNet+AcNet outperform other state-of-the-art models on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3172882, IEEE Transactions on Multimedia



(a) Input (b) Ground truth (c) JORDER PreNet (e) DiNet (Ours) (f) Rain streak (Ours) (g) DiNet+AcNet (Ours)

Fig. 6. Qualitative comparison of different methods for single image rain removal on Rain100H [49] dataset.

TABLE VI
PERFORMANCE OF DIFFERENT MODELS FOR SINGLE IMAGE RAIN REMOVAL ON FOUR DATASETS: **RAIN100L** [49], **TEST2800** [16], **TEST1200** [77], **TEST100** [50], IN TERMS OF PSNR, SSIM, AND FSIM. **HIGHER** VALUE INDICATES BETTER PERFORMANCE.

Methods	Dataset				
	Test100 [50]	**Rain100L** [49]	**Test1200** [77]	**Test2800** [16]	Average
DerainNet [46] (TIP'17)	22.77 / 0.810 / 0.884	27.03 / 0.884 / 0.904	23.38 / 0.835 / 0.924	24.31 / 0.861 / 0.930	24.37 / 0.847 / 0.910
RESCAN [78] (ECCV'18)	25.00 / 0.835 / 0.909	29.80 / 0.881 / 0.919	30.51 / 0.882 / 0.944	31.29 / 0.904 / 0.952	29.14 / 0.874 / 0.930
DIDMDN [77] (CVPR'18)	22.56 / 0.818 / 0.899	25.23 / 0.741 / 0.861	29.65 / 0.901 / 0.950	28.13 / 0.867 / 0.943	26.38 / 0.831 / 0.913
UMRL [79] (CVPR'19)	24.41 / 0.829 / 0.910	29.18 / 0.923 / 0.940	30.55 / 0.910 / 0.955	29.97 / 0.905 / 0.955	28.52 / 0.892 / 0.939
SEMI [80] (CVPR'19)	22.35 / 0.788 / 0.887	25.03 / 0.842 / 0.893	26.05 / 0.822 / 0.917	24.43 / 0.782 / 0.897	24.46 / 0.808 / 0.898
PreNet [6] (CVPR'19)	24.81 / 0.851 / 0.916	32.44 / 0.950 / 0.956	31.36 / 0.911 / 0.955	31.75 / 0.916 / 0.956	30.08 / 0.906 / 0.945
MSPFN [10] (CVPR'20)	27.50 / 0.876 / 0.928	32.40 / 0.933 / 0.943	32.39 / 0.916 / 0.960	32.82 / 0.930 / 0.966	31.27 / 0.913 / 0.948
DiNet (proposed)	27.42 / 0.922 / 0.930	30.40 / 0.959 / 0.958	30.29 / 0.932 / 0.954	31.24 / 0.954 / 0.962	29.83 / 0.941 / 0.951
DiNet + AcNet (proposed)	**27.80 / 0.931 / 0.950**	**34.32 / 0.979 / 0.977**	**32.43 / 0.952 / 0.962**	**33.67 / 0.969 / 0.974**	**32.05 / 0.957 / 0.965**

TABLE VII
DERAINING RESULTS (PSNR / SSIM) OF DIFFERENT METHODS ON RAIN100H [49].

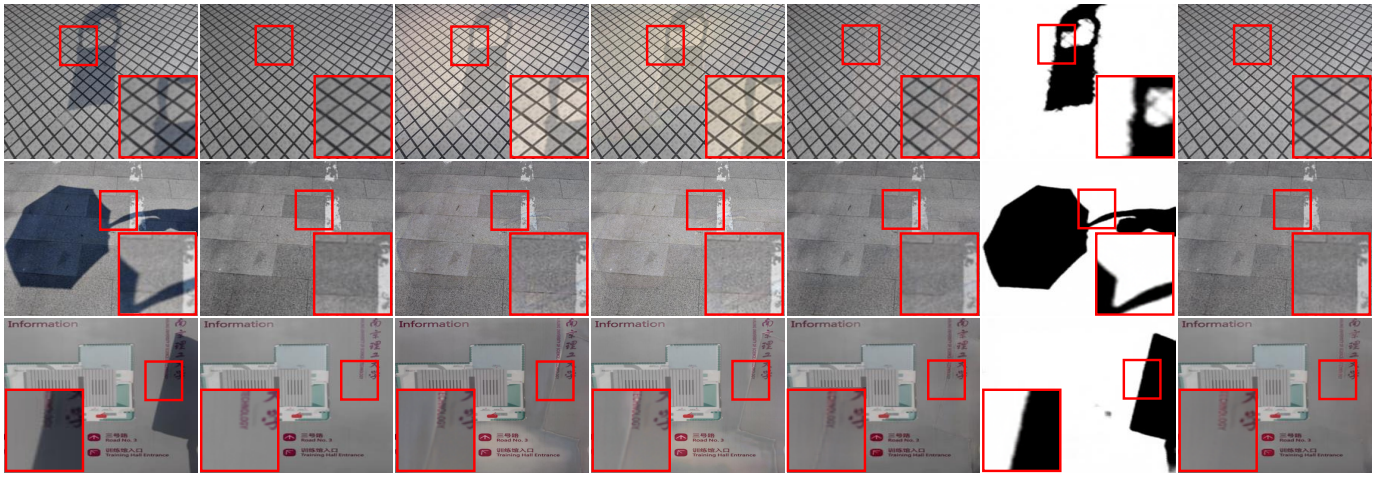
Methods	Rain100H [49]
DDN [16] (CVPR'17)	22.26 / 0.693
JORDER [49] (CVPR'17)	23.45 / 0.749
RESCAN [78] (ECCV'18)	26.45 / 0.846
DIDMDN [77] (CVPR'18)	25.00 / 0.754
DAF-Net [5] (CVPR'19)	28.44 / 0.874
PReNet [6] (CVPR'19)	29.46 / 0.899
DAD [20] (CVPR'20)	30.85 / 0.932
DiNet (proposed)	31.27 / 0.940
DiNet + AcNet (proposed)	**31.82 / 0.946**

both the non-linear mixing dataset [75] and the linear mixing dataset [39]. Moreover, although DiNet alone has achieved better results than other models, with the help of AcNet, the performance of the entire model can be further improved. Table V shows the quantitative results of different models for signle image reflection removal on two real datasets [3], [40]. It manifests that the proposed DiNet and DiNet+AcNet achieve the best performance in most real cases. To gain more insight into the performance of different models on the task of single image reflection removal, we visualize some examples of the results generated by different models in Fig. 4. We notice that our DiNet can remove the reflection more effectively than other three models. Furthermore, with the help of AcNet, the entire model can generate clearer and higher-quality background images.

D. Application: Single Image Rain Removal

Experimental settings. We conduct single image rain removal experiments with two experimental specifications. We first follow the experimental settings in [10] to conduct the large-scale training/testing experiment. The dataset presented by Jiang *et al.* contains 13700 clean/rain image pairs from [16], [50] for training the network. Four test datasets are used to compare the performances of different methods including *Rain100L* [49], *Test2800* [16], *Test1200* [77], and *Test100* [50], which contain 100, 2800, 1200, 100 clean/rain image pairs, respectively. We further validate the effectiveness of our model on a relatively small but difficult dataset, *i.e., Rain100H* (heavy-rain cases) [49], which includes 1800 images for training and 100 images for testing. Since the ground truth rain streaks of the training images in Jiang *et al.* [10] are not available, we simply set the ground truth of the second output as a "zero image" [20] and only train one pathway of the AcNet during training.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3172882, IEEE Transactions on Multimedia



(a) Input (b) Ground truth (c) DSC (d) DSC+ (e)) DiNet (Ours) (f) Shadow mask (Ours) (g) DiNet+AcNet (Ours)

Fig. 7. Qualitative comparison of different methods for single image shadow removal on the ISTD [8] dataset.

Input Ground truth DSC+ DiNet DiNet+AcNet

Fig. 8. Qualitative comparison of different methods for single image shadow removal on the SRD [17] dataset.

We compare our model with state-of-the-art methods for single image reflection removal including: *DDN* [16], *JORDER* [49], *DerainNet* [46], *RESCAN* [78], *DIDMDN* [77], *DAF-Net* [5], *UMRL* [79], *SEMI* [80], *PreNet* [6], and *MSPFN* [10]. All these methods are trained and tested under the same training and testing splits as mentioned above.

Results. Table VI presents the performances of different models for single image rain removal on four test datasets, *i.e.,* Rain100L [49], Test2800 [16], Test1200 [77], and Test100 [50]. It quantitatively manifests that our DiNet and DiNet+AcNet achieve the state-of-the-art performances on the single image rain removal task. Moreover, with the help of AcNet, the performance of the entire model can outperform other state-of-the-art single image rain removal methods. Fig. 5 presents the visualization results of the different methods for single image rain removal. It qualitatively manifests that our DiNet can effectively separate the rain streak from a rainy image but may over-smooth some details. It is worth mentioning that our DiNet+AcNet can not only remove the rain streaks but also keep and restore the image details. Moreover, our methods also achieve the best performance on a relatively difficult dataset: Rain100H, as shown in Table VII. Since the ground truth rain streaks are available in Rain100H, we also show the visualization examples of the generated rain streaks in Fig. 6. It further manifests that our methods can effectively separate rain streaks and clean backgrounds.

TABLE VIII
PERFORMANCE OF DIFFERENT MODELS FOR IMAGE SHADOW REMOVAL ON SRD DATASET [17] AND ISTD DATASET [8] (IN TERMS OF RMSE (**LOWER** IS BETTER)).

Methods	SRD [17]	ISTD [8]
Yang et al. [81] (TIP'12)	22.57	15.63
Guo et al. [53] (TPAMI'12)	12.60	9.300
Gong et al. [82] (BMVC'14)	8.730	8.530
DeshadowNet [17] (CVPR'17)	6.640	7.830
ST-CGAN [8] (CVPR'18)	8.230	7.470
DSC [83] (TPAMI'19)	6.210	6.670
ARGAN [7] (CVPR'19)	5.740	6.680
DAD [20] (CVPR'20)	5.823	6.566
DiNet (proposed)	5.384	6.277
DiNet + AcNet (proposed)	**5.076**	**5.613**

E. Application: Single Image Shadow Removal

Experimental settings. We validate the performance of the proposed method for the single image shadow removal task on two frequently used datasets: *SRD* [17] and *ISTD* [8]. The SRD dataset [17] contains 3088 shadow/shadow-free image pairs, among which 2680 and 408 images are split for training and test respectively. The ISTD dataset [8] contains 1870 shadow/shadow-free/shadow-mask image triplets, among which 1330 and 540 images are split for training and test respectively. We compare our methods with 8 SOTA image shadow removal models, including *DeshadowNet* [17], *ST-CGAN* [8], *DSC* [83], *ARGAN* [7], *DAD* [20], *etc.*

Results. Table VIII presents the quantitative results of different methods for the single image shadow removal. We use the evaluation criterion used by Guo *et al.* [53], *i.e.,* RMSE, of which the lower score is better. It manifests that our proposed methods outperform other methods on both two datasets. Fig. 7 shows the qualitative comparison of our methods with DSC and DSC+ [83] on the ISTD [8] dataset. We notice that DSC and DSC+ can remove the shadow but introduce obvious artifacts in shadow areas. The proposed DiNet can remove the shadow without obvious artifacts. However, there is still

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3172882, IEEE Transactions on Multimedia

TABLE IX
QUANTITATIVE COMPARISON RESULTS ON REAL TEST IMAGES IN *LOL-Real* DATASET [84].

Metric	RRM [56]	SRIE [85]	DRD [84]	DeepUPE [86]	SICE [87]	EG [58]	DRBN [59]	DiNet	AcNet
PSNR	17.34	17.34	15.47	13.27	19.40	18.23	20.13	20.73	**22.16**
SSIM	0.685	0.685	0.567	0.452	0.690	0.616	0.829	0.889	**0.908**

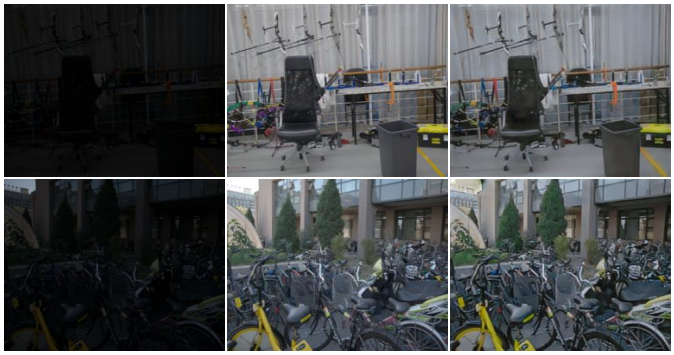


Fig. 9. Visualization of the restored background images by our model for low light enhancement.

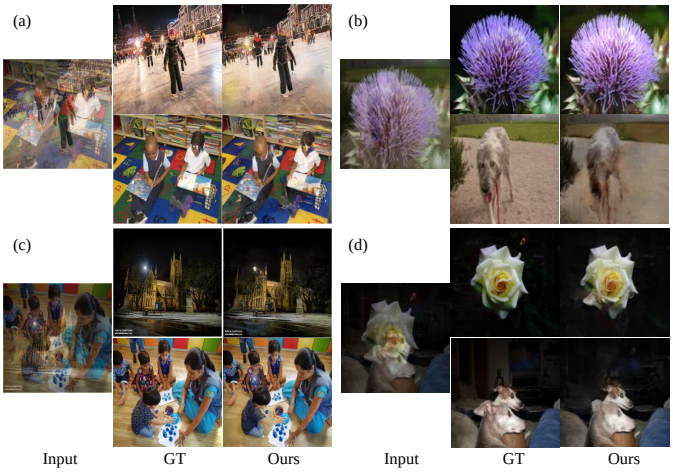


Fig. 10. Failure cases of our method.

some residual information from the shadow. After the process of AcNet, we observe that the residual information from the shadow is less than before, and the perceptual quality is better. Moreover, the shadow masks shown in Fig. 7 also demonstrate that our method can effectively separate the shadow masks and the background images. Fig. 8 shows the qualitative comparison of our methods with DSC+ [83] on the SRD [8] dataset, which further validates the effectiveness of our methods.

F. Application: Illumination Correction

Experimental settings. We further conduct the experiment for one single image illumination correction task, *i.e.,* low light enhancement, on *LOL-Real* dataset [84]. We follow the experimental settings of the supervised stage in [59] to conduct the experiment, which splits the LOL dataset into 689 and 100 low-light/normal-light image pairs for training and testing, respectively. Seven state-of-the-art methods including *RRM* [56], *SRIE* [85], *DRD* [84], *DeepUPE* [86], *SICE* [87], *EG* [58], and *DRBN* [59] are introduced for comparison.

Results. Table IX presents the performances of different models for low-light enhancement on LOL-Real dataset [84]. It quantitatively manifests the superiority of our methods. It is worth mentioning that even though DRBN [59] uses additional datasets for two-stage semi-supervised training, our method still performs better than DRBN. Fig. 9 also manifests that our method can effectively enhance the low-light images.

G. Failure Cases

As shown in Fig. 10, our method also has some failure cases. It can be observed from the first two examples that if the superimposed parts are produced by the superimposition of two complex textures, such as faces, the restored results are not very good. Moreover, if one image layer is strongly suppressed by another layer, the restored results are also not satisfactory. Fig. 10 (c) (d) show two example results of our method for low-light cases. It can be observed that for normal light contents, our method can still perform well, while for low-light area, the separation results may not be good.

V. CONCLUSION

In this paper, we define the superimposed image separation problem as the visual confusion decomposition task which aims to restore two separated images and eliminate the visual confusion from the non-target image. Inspired by the development then rivalry process of human vision on a single superimposed image, we propose a framework for single superimposed image decomposition, which mainly includes a differentiation net, an activation net, and multiple loss functions. Experimental results indicate that the proposed framework achieves the state-of-the-art performance on the superimposed image separation task and multiple related applications, including single image reflection removal, single image de-raining, single image shadow removal, *etc*.

Our approach has great potential since it can be used for many image decomposition tasks, and can achieve good results. Moreover, besides being used for image decomposition, our method may also contribute to other signal decomposition tasks, such as audio separation. However, like most two/multi-stage methods, the training/inference procedures of our method are more complex compared to one-stage methods. It is significant to consider integrating our two networks into a one-stage. Moreover, our methods need to be trained on different tasks respectively to work fine. Our future works will explore how to train our model on mixed tasks to handle all these tasks simultaneously.

REFERENCES

[1] E. Peli and J.-H. Jung, "Multiplexing prisms for field expansion," *Optometry and Vision Science: Official Publication of the American Academy of Optometry*, vol. 94, no. 8, p. 817, 2017.

[2] H. Duan, X. Min, Y. Zhu, G. Zhai, X. Yang, and P. L. Callet, "Confusing image quality assessment: Towards better augmented reality experience," *arXiv preprint arXiv:2204.04900*, 2022.

[3] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3922–3930.

[4] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang, "Single image reflection removal exploiting misaligned training data and network enhancements," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8178–8187.

[5] X. Hu, C.-W. Fu, L. Zhu, and P.-A. Heng, "Depth-attentional features for single-image rain removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8022–8031.

[6] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, "Progressive image deraining networks: A better and simpler baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3937–3946.

[7] B. Ding, C. Long, L. Zhang, and C. Xiao, "Argan: Attentive recurrent generative adversarial network for shadow detection and removal," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 10 213–10 222.

[8] J. Wang, X. Li, and J. Yang, "Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1788–1797.

[9] Z. Deng, L. Zhu, X. Hu, C.-W. Fu, X. Xu, Q. Zhang, J. Qin, and P.-A. Heng, "Deep multi-model fusion for single-image dehazing," in *in Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 2453–2462.

[10] K. Jiang, Z. Wang, P. Yi, C. Chen, B. Huang, Y. Luo, J. Ma, and J. Jiang, "Multi-scale progressive fusion network for single image deraining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8346–8355.

[11] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 29, no. 9, pp. 1647–1654, 2007.

[12] K. Gai, Z. Shi, and C. Zhang, "Blindly separating mixtures of multiple layers with spatial shifts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2008, pp. 1–8.

[13] K. Gai, Z. Shi, and C. Zhang, "Blind separation of superimposed images with unknown motions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 1881–1888.

[14] K. Gai, Z. Shi, and C. Zhang, "Blind separation of superimposed moving images using image statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 1, pp. 19–32, 2011.

[15] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3238–3247.

[16] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3855–3863.

[17] L. Qu, J. Tian, S. He, Y. Tang, and R. W. Lau, "Deshadownet: A multi-context embedding deep network for shadow removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4067–4075.

[18] C. Li, Y. Yang, K. He, S. Lin, and J. E. Hopcroft, "Single image reflection removal through cascaded refinement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3565–3574.

[19] Y. Gandelsman, A. Shocher, and M. Irani, "double-dip": Unsupervised image decomposition via coupled deep-image-priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 6, 2019, p. 2.

[20] Z. Zou, S. Lei, T. Shi, Z. Shi, and J. Ye, "Deep adversarial decomposition: A unified framework for separating superimposed images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 806–12 816.

[21] R. Blake and N. K. Logothetis, "Visual competition," *Nature Reviews Neuroscience*, vol. 3, no. 1, pp. 13–21, 2002.

[22] R. P. O'Shea, A. Parker, D. La Rooy, and D. Alais, "Monocular rivalry exhibits three hallmarks of binocular rivalry: Evidence for common processes," *Vision Research*, vol. 49, no. 7, pp. 671–681, 2009.

[23] H. Duan, W. Shen, X. Min, D. Tu, J. Li, and G. Zhai, "Saliency in augmented reality," *arXiv preprint arXiv:2204.08308*, 2022.

[24] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7151–7160.

[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[26] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2482–2491.

[27] H. Apfelbaum and E. Peli, "Tunnel vision prismatic field expansion: challenges and requirements," *Translational Vision Science & Technology*, vol. 4, no. 6, pp. 8–8, 2015.

[28] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, no. 11, pp. 1–52, 2020.

[29] G. Zhai, W. Zhang, X. Yang, and Y. Xu, "Image quality assessment metrics based on multi-scale edge presentation," in *IEEE Workshop on Signal Processing Systems Design and Implementation*, 2005, pp. 331–336.

[30] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. Le Callet, "How is gaze influenced by image transformations? dataset and model," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 2287–2300, 2019.

[31] X. Min, G. Zhai, J. Zhou, M. C. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 6054–6068, 2020.

[32] J. Xu, W. Zhou, H. Li, F. Li, and Q. Jiang, "Quality assessment of multi-exposure image fusion by synthesizing local and global intermediate references," *Displays*, vol. 74, p. 102188, 2022.

[33] B. Wolfe, J. Dobres, R. Rosenholtz, and B. Reimer, "More than the useful field: Considering peripheral vision in driving," *Applied Ergonomics*, vol. 65, pp. 316–325, 2017.

[34] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[35] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2752–2759.

[36] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection removal using ghosting cues," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3193–3201.

[37] N. Arvanitopoulos, R. Achanta, and S. Susstrunk, "Single image reflection suppression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4498–4506.

[38] H. Zhang, X. Xu, H. He, S. He, G. Han, J. Qin, and D. Wu, "Fast user-guided single image reflection removal via edge-aware cascaded networks," *IEEE Transactions on Multimedia (TMM)*, vol. 22, no. 8, pp. 2012–2023, 2019.

[39] J. Yang, D. Gong, L. Liu, and Q. Shi, "Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 654–669.

[40] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4786–4794.

[41] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 694–711.

[42] Y. Li, M. Liu, Y. Yi, Q. Li, D. Ren, and W. Zuo, "Two-stage single image reflection removal with reflection-aware guidance," *arXiv preprint arXiv:2012.00945*, 2020.

[43] W. Yang, R. T. Tan, S. Wang, Y. Fang, and J. Liu, "Single image deraining: From model-based to data-driven and beyond," *IEEE Transactions*

on Pattern Analysis and Machine Intelligence (TPAMI), vol. 43, no. 11, pp. 4059–4077, 2020.

[44] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown, "Rain streak removal using layer priors," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 2736–2744.

[45] Y. Wang, S. Liu, C. Chen, and B. Zeng, "A hierarchical approach for rain or snow removing in a single color image," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 8, pp. 3936–3950, 2017.

[46] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: A deep network architecture for single-image rain removal," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 6, pp. 2944–2956, 2017.

[47] Y. Wang, D. Gong, J. Yang, Q. Shi, A. van den Hengel, D. Xie, and B. Zeng, "Deep single image deraining via modeling haze-like effect," *IEEE Transactions on Multimedia (TMM)*, 2020.

[48] X. Lin, L. Ma, B. Sheng, Z.-J. Wang, and W. Chen, "Utilizing two-phase processing with fbls for single image deraining," *IEEE Transactions on Multimedia (TMM)*, vol. 23, pp. 664–676, 2020.

[49] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1357–1366.

[50] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2019.

[51] G. D. Finlayson, M. S. Drew, and C. Lu, "Entropy minimization for shadow removal," *International Journal of Computer Vision (IJCV)*, vol. 85, no. 1, pp. 35–57, 2009.

[52] X. Huang, G. Hua, J. Tumblin, and L. Williams, "What characterizes a shadow boundary under the sun and sky?" in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2011, pp. 898–905.

[53] R. Guo, Q. Dai, and D. Hoiem, "Paired regions for shadow detection and removal," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 12, pp. 2956–2967, 2012.

[54] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

[55] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, "A dynamic histogram equalization for image contrast enhancement," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, pp. 593–600, 2007.

[56] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 6, pp. 2828–2841, 2018.

[57] S. Hao, X. Han, Y. Guo, X. Xu, and M. Wang, "Low-light image enhancement with semi-decoupled decomposition," *IEEE Transactions on Multimedia (TMM)*, vol. 22, no. 12, pp. 3025–3038, 2020.

[58] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 2340–2349, 2021.

[59] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3063–3072.

[60] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 3461–3473, 2021.

[61] X. Feng, W. Pei, Z. Jia, F. Chen, D. Zhang, and G. Lu, "Deep-masking generative network: A unified framework for background restoration from superimposed images," *IEEE Transactions on Image Processing (TIP)*, vol. 30, pp. 4867–4882, 2021.

[62] C. Wu, J. Zhang, and X.-C. Tai, "Augmented lagrangian method for total variation restoration with non-quadratic fidelity," *Inverse Problems & Imaging*, vol. 5, no. 1, p. 237, 2011.

[63] C. Wu, J. Zhang, Y. Duan, and X.-C. Tai, "Augmented lagrangian method for total variation based image restoration and segmentation over triangulated surfaces," *Journal of Scientific Computing*, vol. 50, no. 1, pp. 145–166, 2012.

[64] Z. Pan, F. Yuan, J. Lei, Y. Fang, X. Shao, and S. Kwong, "Vcrnet: Visual compensation restoration network for no-reference image quality assessment," *IEEE Transactions on Image Processing (TIP)*, 2022.

[65] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[66] Y. Wu and K. He, "Group normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[67] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7794–7803.

[68] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[69] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[70] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 447–456.

[71] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (CVPR)*, vol. 2, no. 1, 2011.

[72] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 1447–1454.

[73] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[74] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[75] Q. Wen, Y. Tan, J. Qin, W. Liu, G. Han, and S. He, "Single image reflection removal beyond linearity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3771–3779.

[76] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232.

[77] H. Zhang and V. M. Patel, "Density-aware single image de-raining using a multi-stream dense network," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 695–704.

[78] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 254–269.

[79] R. Yasarla and V. M. Patel, "Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8405–8414.

[80] W. Wei, D. Meng, Q. Zhao, Z. Xu, and Y. Wu, "Semi-supervised transfer learning for image rain removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3877–3886.

[81] Q. Yang, K.-H. Tan, and N. Ahuja, "Shadow removal using bilateral filtering," *IEEE Transactions on Image processing (TIP)*, vol. 21, no. 10, pp. 4361–4368, 2012.

[82] H. Gong and D. Cosker, "Interactive shadow removal and ground truth for variable scene categories." in *BMVC*, 2014, pp. 1–11.

[83] X. Hu, C. Fu, L. Zhu, J. Qin, and P. Heng, "Direction-aware spatial context features for shadow detection and removal." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[84] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," *arXiv preprint arXiv:1808.04560*, 2018.

[85] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2782–2790.

[86] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6849–6857.

[87] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 4, pp. 2049–2062, 2018.

Huiyu Duan received the B.E. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. From Sept. 2019 to Sept. 2020, he was a visiting Ph.D. student at the Schepens Eye Research Institute, Harvard Medical School, Boston, USA. His research interests include perceptual quality assessment, quality of experience, visual attention modeling, extended reality (XR), and multimedia signal processing.

Wei Shen is a tenure-track Associate Professor at the Artificial Intelligence Institute, Shanghai Jiao Tong University, since October 2020. Before that, he was an Assistant Research Professor at the Department of Computer Science, Johns Hopkins University. His research interests lie in the fields of computer vision, machine learning, and deep learning, particularly in object detection and segmentation, representation learning, human-centered computer vision and medical image analysis. He is an area chair for CVPR 2022 and ACCV 2022, a senior program committee member for AAAI 2022 and an associate editor for Neurocomputing.

Xiongkuo Min received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2018, where he is currently a tenure-track Associate Professor with the Institute of Image Communication and Network Engineering. From Jan. 2016 to Jan. 2017, he was a visiting student at University of Waterloo. From Jun. 2018 to Sept. 2021, he was a Postdoc at Shanghai Jiao Tong University. From Jan. 2019 to Jan. 2021, he was a visiting Postdoc at The University of Texas at Austin. He received the Best Paper Runner-up Award of IEEE Transactions on Multimedia in 2021, the Best Student Paper Award of IEEE International Conference on Multimedia and Expo (ICME) in 2016, and the excellent Ph.D. thesis award from the Chinese Institute of Electronics (CIE) in 2020. His research interests include image/video/audio quality assessment, quality of experience, visual attention modeling, extended reality, and multimodal signal processing.

Yuan Tian received the B.Sc. degree in electronic engineering from Wuhan University, Wuhan, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His works have been published in top-tier journals and conferences (e.g., IJCV, ICCV, and ECCV). His research interests include video understanding and video compression.

Jae-Hyun Jung is an Assistant Scientist at Schepens Eye Research Institute/Massachusetts Eye and Ear, Department of Ophthalmology, Harvard Medical School. He received his Ph.D. in Electrical Engineering and Computer Science from Seoul National University, Korea, in 2012. Dr. Jung is a Fellow of the American Academy of Optometry. He was presented the Alice Adler Fellowship Award and Best Paper of the year award by Harvard Medical School, and the Merck Young Scientists Award by the Society for Information Display. Dr. Jung's research focuses on the interface between electrical/optical systems and human vision. He proposes to uncover the principles underlying human visual perception and apply them to improve displays and vision aids using engineering approaches such as AR/VR displays and light-field imaging systems.

Xiaokang Yang (M'00-SM'04-F'19) received the B.S. degree from Xiamen University, Xiamen, China, in 1994, the M.S. degree from the Chinese Academy of Sciences, Shanghai, China, in 1997, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, in 2000. From September 2000 to March 2002, he worked as a Research Fellow with the Centre for Signal Processing, Nanyang Technological University, Singapore. From April 2002 to October 2004, he was a Research Scientist with the Institute for Infocomm Research (I2R), Singapore. From August 2007 to July 2008, he visited the Institute for Computer Science, University of Freiburg, Germany, as an Alexander von Humboldt Research Fellow. He is currently a Distinguished Professor with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. He has published over 200 refereed articles and has filed 60 patents. His current research interests include image processing and communication, computer vision, and machine learning. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and a Senior Associate Editor of the IEEE SIGNAL PROCESSING LETTERS.

Guangtao Zhai (SM'19) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009, where he is currently a Research Professor with the Institute of Image Communication and Information Processing. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Post-Doctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University of Erlangen-Nuremberg, Germany. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012. His research interests include multimedia signal processing and perceptual signal processing.