A UNIFIED TWO-STAGE MODEL FOR SEPARATING SUPERIMPOSED IMAGES

Huiyu Duan, Xiongkuo Min, Wei Shen, and Guangtao Zhai

{huiyuduan, minxiongkuo, wei.shen, zhaiguangtao}@sjtu.edu.cn
Shanghai Jiao Tong University, Shanghai, China

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. In this paper, we propose a human vision-inspired framework for separating superimposed images. 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.

Index Terms— Superimposed image decomposition, develop then rival, two stage, reflection removal, rain removal.

1. INTRODUCTION

Visual confusion [1] (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 [2, 3], image de-raining [4, 5], *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*. 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 [6], 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 [6]. 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 [4, 7–10]. Nevertheless, most of them only focused on one specific separation case, while a unified framework is rarely considered. Gandelsman *et al.* [11] 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.* [12] 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 [13, 14] 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 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 two-stage framework for single superimposed image separation. 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, contextual attention module [15, 16] is integrated into the network. A multi-scale [17] crossroad perceptual loss is introduced, 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. 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. 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.

2. 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 on 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,

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

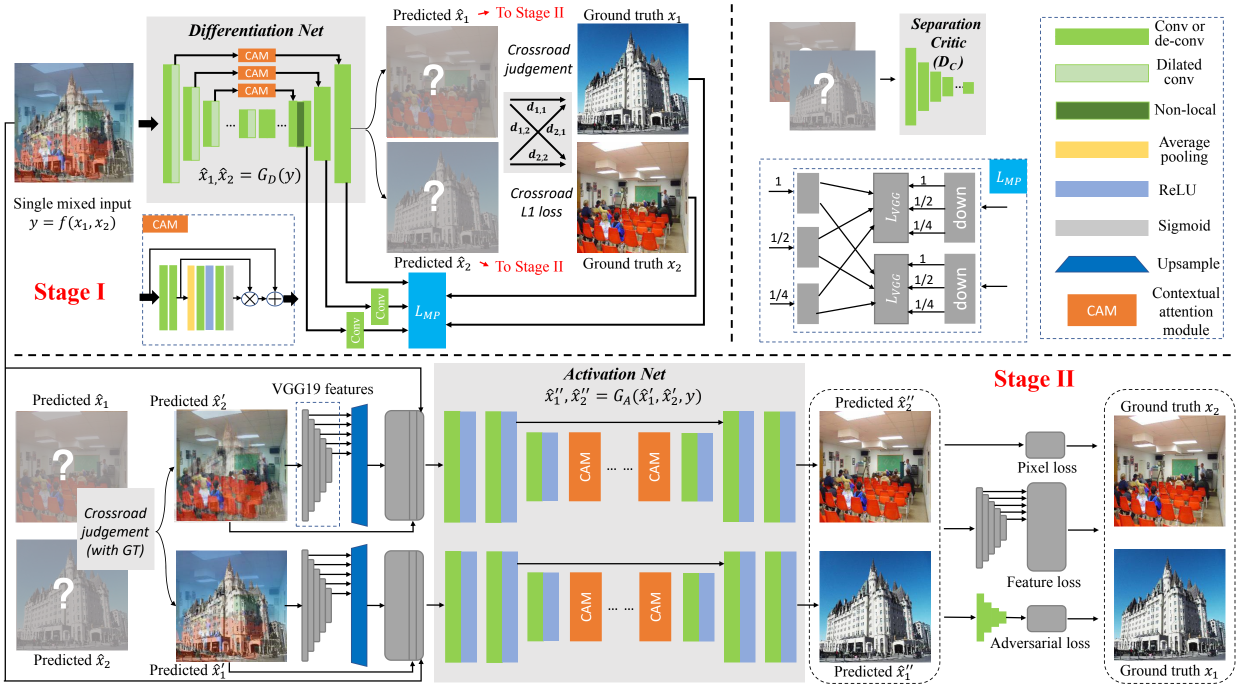

Fig. 1: An overview of the proposed method. The proposed method consists of two stages, including "Stage I: development stage" and "Stage II: rivalry stage".

which tries to activate one layer and suppress another layer. Fig. 1 demonstrates the overview of the proposed framework.

2.1. 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).

2.1.1. 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" [12,18]. 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 in the decoder part to better perceive the whole image. We also leverages the contextual attention module [16] (illustrated as CAM in Fig. 1, *a.k.a*, channel attention module) to introduce global contextual information across channels for better disentangling superimposed features.

2.1.2. Objective Function

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

Crossroad $\mathcal{L}_1$ Loss. Since the order of the decomposition outputs is not specified, we use the crossroad $\mathcal{L}_1$ loss [12] to measure the pixel-wise distance between the predicted outputs and the ground truths, which is defined as:

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

where $d_{i,j} = \|\hat{x}_i - x_j\|, i, j \in \{1, 2\}$.

Separation Critic. To further improve the separation performance, a decomposition prior learned through an adversarial training is introduced [12], 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:

$$\begin{aligned}\mathcal{L}^{D_C}_{\text{critic}} &= \mathbb{E}_{x_i \sim p_i(x_i)}\{\log D_C(x_1, x_2)\} \\ &+ \mathbb{E}_{\hat{x}_i \sim p_i(\hat{x}_i)}\{\log(1 - D_C(\hat{x}_1, \hat{x}_2))\},\end{aligned} \tag{3}$$

where $D_C(x, y)$ is the probability that the pair (x, y) is a well-separated (clean) image pair. The loss function of the generator G_D is defined as:

$$\mathcal{L}^{G_D}_{\text{critic}} = \mathbb{E}_{\hat{x}_i \sim p_i(\hat{x}_i)}\{-\log(D_C(\hat{x}_1, \hat{x}_2))\}. \tag{4}$$

Multi-scale Crossroad Perceptual Loss. Multi-scale losses are proved to be effective in optimizing image decomposition tasks such as de-raining [17] and reflection removal. 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. We adopt the perceptual distance over different scales rather than other loss functions in order to utilize both low-level and high-level information. Since the order of the decomposition outputs of multiple scales is also not specified, 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:

$$\begin{aligned}\hat{x}'_1 = \hat{x}_i&, \hat{x}'_2 = \hat{x}_j, \\ \textbf{s.t.}\ \min&\{d_{i,1} + d_{j,2}\},\end{aligned} \tag{5}$$

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:

$$\mathcal{L}_{\text{MP}} = \sum_{k=1}^{M}(\lambda_k(\mathcal{L}_{\text{VGG}}(\hat{x}'_{1_k}, x_{1_k}) + \mathcal{L}_{\text{VGG}}(\hat{x}'_{2_k}, x_{2_k}))), \tag{6}$$

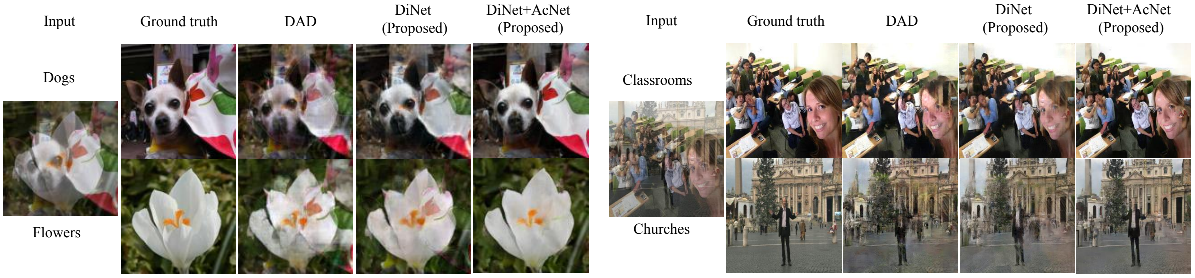

Fig. 2: Qualitative comparison of superimposed image decomposition on two mixing datasets.

Table 1: Performance (PSNR / SSIM) of different methods for superimposed image separation on two mixing datasets: 1) Dogs [20] + Flwrs [21], and 2) LSUN Classroom + LSUN Church [22]. (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.* [6] (TPAMI'07)	10.54 / 0.444	10.46 / 0.366
Double-DIP [11] (CVPR'19)	14.70 / 0.661	13.83 / 0.590
DAD [12] (CVPR'20)	25.51 / 0.849	26.32 / 0.883
DiNet	26.65 / 0.876	27.13 / 0.901
DiNet + AcNet	**28.82 / 0.918**	**29.88 / 0.940**

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. $\mathcal{L}_{\text{VGG}}$ is the well-known perceptual (feature) loss function [19].

Our overall loss function of DiNet is:

$$\mathcal{L}_{\text{di}} = \alpha_1 \mathcal{L}_{\text{cross}} + \alpha_2 \mathcal{L}_{\text{critic}} + \alpha_3 \mathcal{L}_{\text{MP}}, \tag{7}$$

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.

2.2. 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 cross-road judgement module as described in Section 2.1.2 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:

$$\hat{x}''_1, \hat{x}''_2 = G_A(\hat{x}'_1, \hat{x}'_2, y), \tag{8}$$

where G_A is the proposed activation net (AcNet).

2.2.1. *Network Architecture*

The architecture of the proposed activation network (AcNet) is built based on the Resnet generator [19] 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 [23] from a pretrained VGG-19 network [24], and then concatenate these features with $\hat{x}'_1$ or $\hat{x}'_2$, and the single mixed input y as an augmented network input. The AcNet contains 7 cascaded CAM blocks of which the architecture is the same with that in "Stage I".

2.2.2. *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:

$$\mathcal{L}_{\text{pixel}} = \mathbb{E}_{(\hat{x}''_i, x_i) \sim p_i(\hat{x}''_i, x_i)} \{\mathcal{L}_1(\hat{x}''_i, x_i)\}. \tag{9}$$

Table 2: Ablation studies for the architecture and losses of DiNet on two mixing datasets: 1) Dogs [20] + Flwrs [21], and 2) LSUN Classroom + LSUN Church [22].

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**

Table 3: Ablation studies for the architecture and losses of AcNet on the Dogs [20] + Flwrs [21] dataset.

Methods	w/o MI	3CAM	w/o $\mathcal{L}_{\text{adv}}$	w/o $\mathcal{L}_{\text{feat}}$	all combined
PSNR	27.34	27.63	28.92	27.92	**28.82**
SSIM	0.890	0.901	0.920	0.903	**0.918**

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:

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

where $\mathcal{L}_{\text{VGG}}$ is the same as that mentioned in Section 2.1.2.

Adversarial Loss. To encourage the predicted output to be as realistic as the ground-truth image layer, an adversarial loss [25] is used to improve the realism of the predicted output. The loss function of the discriminator D is defined as:

$$\begin{aligned} \mathcal{L}^D_{\text{adv}} = \mathbb{E}_{(\hat{x}''_i, y_i) \sim p_i(\hat{x}''_i, y_i)} \{\log D(\hat{x}''_i, y_i)\} \\ - \mathbb{E}_{(x_i, y_i) \sim p_i(x_i, y_i)} \{\log D(x_i, y_i)\}, \end{aligned} \tag{11}$$

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

$$\mathcal{L}^G_{\text{adv}} = -\mathbb{E}_{(\hat{x}''_i, y_i) \sim p_i(\hat{x}''_i, y_i)} \{\log D(\hat{x}''_i, y_i)\}. \tag{12}$$

Our overall loss function for AcNet is:

$$\mathcal{L}_{\text{ac}} = \beta_1 \mathcal{L}_{\text{pixel}} + \beta_2 \mathcal{L}_{\text{feat}} + \beta_3 \mathcal{L}_{\text{adv}}, \tag{13}$$

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

3. EXPERIMENTAL VALIDATION

We evaluate the proposed method on 3 tasks, including 1) superimposed image separation, 2) single image reflection removal, 3) single image rain removal. The experimental settings and results are described and analyzed in detail as follows.

3.1. Separating Superimposed Images

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 [12] and evaluate the proposed method on two datasets of mixed image decomposition: 1) Stanford-Dogs (Dogs) [20] + VGG-Flowers (Flwrs) [21], and 2)

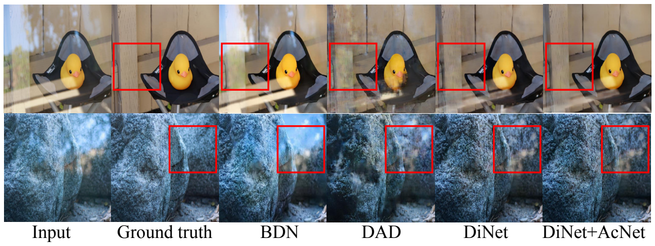

Fig. 3: Qualitative comparison of different methods for reflection removal on two test datasets.

Table 4: Quantitative results (PSNR / SSIM) of different methods for reflection removal on a real dataset [26].

Methods	Real20 [26]	Wild [2]
CEILNet [8] (ICCV'17)	19.04 / 0.762	22.14 / 0.819
Zhang *et al.* [26] (CVPR'18)	21.30 / 0.821	21.52 / 0.829
BDN [27] (ECCV'18)	20.06 / 0.738	22.34 / 0.821
ERRNet [3] (CVPR'19)	22.80 / 0.803	24.16 / 0.847
DAD [12] (CVPR'20)	22.36 / 0.846	24.80 / 0.922
DMGN [28] (TIP'21)	23.05 / 0.823	25.18 / 0.894
DiNet (proposed)	23.11 / 0.870	25.56 / 0.926
DiNet + AcNet (proposed)	**23.80 / 0.877**	**25.69 / 0.929**

LSUN Classroom + LSUN Church [22]. We follow the experimental settings in [12] to conduct the experiments. Table 1 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 models achieve the best performance in terms of both metrics on two datasets. 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 and AcNet separate the superimposed image better compared to *DAD* with less artifacts.

3.2. 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, 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 2 shows the results of the ablation experiments for DiNet, which indicates all components together contribute to the final performance.

Ablation studies for AcNet. We then perform ablation experiments on five variants of the AcNet and present the results in Table 3. We observe significant improvement by incorporating the mixed input to activate one layer rather than only refining the output from the last stage, and less CAM number may decrease the performance. We also analyze the contribution of each loss function in the AcNet, and we notice that the perceptual loss $\mathcal{L}_{\text{feat}}$ contributes the most to the improvement.

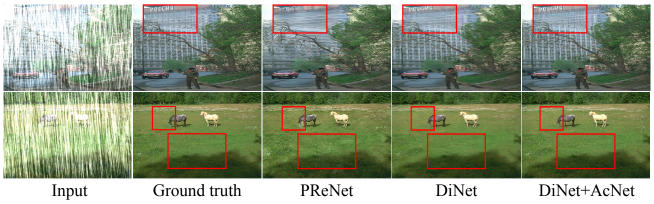

Fig. 4: Qualitative comparison of different methods for single image rain removal task.

Table 5: Deraining results (PSNR / SSIM) of different methods on Rain100H [29] and Rain800 [30].

Methods	Rain100H [29]	Rain800 [30]
DDN [9] (CVPR'17)	22.26 / 0.693	21.16 / 0.732
JORDER [29] (CVPR'17)	23.45 / 0.749	22.29 / 0.792
RESCAN [31] (ECCV'18)	26.45 / 0.846	24.09 / 0.841
PReNet [5] (CVPR'19)	29.46 / 0.899	- / -
DAD [12] (CVPR'20)	30.85 / 0.932	24.49 / 0.885
DiNet (proposed)	31.27 / 0.940	24.50 / 0.885
DiNet + AcNet (proposed)	**31.82 / 0.946**	**25.04 / 0.896**

3.3. Application: Single Image Reflection Removal

We conduct single image reflection removal experiments on two real test datasets [2, 26]. We follow the common training/test methods which have been widely used in the literature [3, 12, 26, 28] 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. 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" [12] and only train one pathway of the AcNet during training. Table 4 shows the quantitative results of different models for signle image reflection removal on two real datasets [2, 26]. 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. 3. 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.

3.4. Application: Single Image Rain Removal

We also conduct experiments for image de-raining on two datasets: Rain100H [29], Rain800 [30]. We follow the training/testing split in the original datasets of Rain100H [29] and Rain800 [30] to conduct the experiment, respectively. We report the performance of the proposed methods and other methods in Table 5. Our methods outperform other SOTA methods in most entries. Figure 4 shows the visual results for image de-raining. It can be seen that the DiNet and AcNet can significantly reduce the artifacts after de-raining.

4. CONCLUSION

In this paper, inspired by the development then rivalry process of human vision on a single superimposed image, we propose a unified two-stage framework for separating superimposed images, 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, *etc.*

5. REFERENCES

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

[2] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot, "Benchmarking single-image reflection removal algorithms," in *Proc. ICCV*, 2017, pp. 3922–3930.

[3] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang, "Single image reflection removal exploiting misaligned training data and network enhancements," in *Proc. CVPR*, 2019, pp. 8178–8187.

[4] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng, "Depth-attentional features for single-image rain removal," in *Proc. CVPR*, 2019, pp. 8022–8031.

[5] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng, "Progressive image deraining networks: A better and simpler baseline," in *Proc. CVPR*, 2019, pp. 3937–3946.

[6] Anat Levin and Yair 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.

[7] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao, "Argan: Attentive recurrent generative adversarial network for shadow detection and removal," in *Proc. ICCV*, 2019, pp. 10213–10222.

[8] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proc. ICCV*, 2017, pp. 3238–3247.

[9] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley, "Removing rain from single images via a deep detail network," in *Proc. CVPR*, 2017, pp. 3855–3863.

[10] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau, "Deshadownet: A multi-context embedding deep network for shadow removal," in *Proc. CVPR*, 2017, pp. 4067–4075.

[11] Yossi Gandelsman, Assaf Shocher, and Michal Irani, "double-dip": Unsupervised image decomposition via coupled deep-image-priors," in *Proc. CVPR*, 2019, vol. 6, p. 2.

[12] Zhengxia Zou, Sen Lei, Tianyang Shi, Zhenwei Shi, and Jieping Ye, "Deep adversarial decomposition: A unified framework for separating superimposed images," in *Proc. CVPR*, 2020, pp. 12806–12816.

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

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

[15] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal, "Context encoding for semantic segmentation," in *Proc. CVPR*, 2018, pp. 7151–7160.

[16] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, 2018, pp. 7132–7141.

[17] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proc. CVPR*, 2018, pp. 2482–2491.

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

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*. Springer, 2016, pp. 694–711.

[20] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR*, 2011, vol. 2.

[21] M-E Nilsback and Andrew Zisserman, "A visual vocabulary for flower classification," in *Proc. CVPR*, 2006, vol. 2, pp. 1447–1454.

[22] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[23] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. CVPR*, 2015, pp. 447–456.

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

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 1125–1134.

[26] Xuaner Zhang, Ren Ng, and Qifeng Chen, "Single image reflection separation with perceptual losses," in *Proc. CVPR*, 2018, pp. 4786–4794.

[27] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi, "Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal," in *Proc. ECCV*, 2018, pp. 654–669.

[28] Xin Feng, Wenjie Pei, Zihui Jia, Fanglin Chen, David Zhang, and Guangming 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.

[29] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan, "Deep joint rain detection and removal from a single image," in *Proc. CVPR*, 2017, pp. 1357–1366.

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

[31] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Proc. ECCV*, 2018, pp. 254–269.