Variational Degeneration to Structural Refinement: A Unified Framework for Superimposed Image Decomposition

Wenyu Li[1], Yan Xu[1], Yang Yang[1]*, Haoran Ji[1], Yue Lang[2]
[1]Tianjin University, Tianjin, China.
[2]Hebei University of Technology, Tianjin, China.

Abstract

Decomposing a single mixed image into individual image layers is the common crux of a classical category of tasks in image restoration. Several unified frameworks have been proposed that can handle different types of degradation in superimposed image decomposition. However, there are always undesired structural distortions in the separated images when dealing with complicated degradation patterns. In this paper, we propose a unified framework for superimposed image decomposition that can cope with intricate degradation patterns adaptively. Considering the different mixing patterns between the layers, we introduce a degeneration representation in the latent space to mine the intrinsic relationship between the superimposed image and the degeneration pattern. Moreover, by extracting structure-guided knowledge from the superimposed image, we further propose structural guidance refinement to avoid confusing content caused by structure distortion. Extensive experiments have demonstrated that our method remarkably outperforms other popular image separation frameworks. The method also achieves competitive results on related applications including image deraining, image reflection removal, and image shadow removal, which validates the generalization of the framework.

1. Introduction

Single superimposed image decomposition aims to decompose a given superimposed image into the corresponding source images. It involves many critical research tasks, such as image deraining, reflection removal, and shadow removal, etc. The key feature of this type of task is that the input degraded image can be viewed as a superimposing of two layers. For example, image deraining can be treated as decomposing a rainy image into a rain-free image and rain streaks.

In superimposed image decomposition, the degradation model is formulated as follows:

$$I = g\left(x_1\right) + f\left(x_2\right), \tag{1}$$

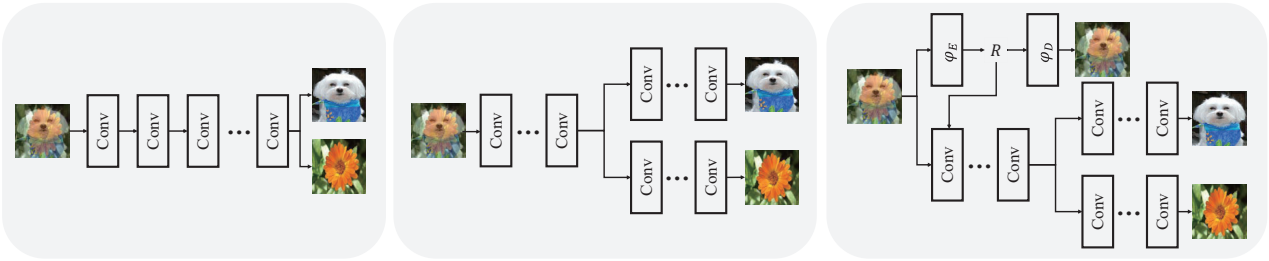
where $g\left(\cdot\right)$ and $f\left(\cdot\right)$ represent various degradation representations for x_1 and x_2, which act as crucial components to model the degradation. A challenge of this task is dealing with various and complicated degradation patterns produced in different degradation processes, like the mixing factor of two layers in Eq. (1). These degradation patterns are extremely difficult to identify, therefore, previous methods usually formulate these tasks as individual research problems, making great progress [8, 42, 15]. However, the model well-designed for one task is hard to apply directly to another, due to the degradation changes. There is a high expectation to handle all the above tasks within a unified framework.

In recent years, existing studies have exploited the Unet-based [33] single lane structure to restore source images in a unified framework (see Figure 2). While high-fidelity results can be generated, this structure always suffers from

(a) Input (b) DAD (c) GIP
(d) BIDeN (e) VDSR (Ours) (f) Ground truth

Figure 1. Image decomposition results on the Stanford-Dogs [23] + VGG-Flowers [29] dataset. We only present one output here. It can be observed that both DAD [51] and GIP [52] contain blurred edges of "flower", and BIDeN [14] produces slight blur and color distortion. Our result preserve finer geometric structures.

*Corresponding author.



(a) Single lane structure (b) Multi-lane structure (c) Ours

Figure 2. Image decomposition paradigms. (a) Single lane structure with single encoder and decoder. (b) Multi-lane structure with shared encoder and multiple decoders. (c) our proposed variational degradation module to jointly learn degeneration representation R and image decomposition.

structural inconsistency including geometric distortions and texture copy. As Figure 1 shows, one layer contains artifacts and residuals from another layer, since the single lane may introduce unstable factors and is insufficient to handle complicated degradation patterns [21]. Subsequently, a multi-lane structure with two branches can split the decomposition problem into sub-problems [30]. The structure performs the decomposition via the explicitly separated decoders. However, the intrinsic problem is not alleviated, since existing models neglect to model the degeneration representations and directly complete the decomposition task, which fails to preserve structural information in the complex degradation patterns.

In this paper, we propose a novel unified framework for single superimposed image decomposition, which is named as Variational Degeneration to Structural Refinement (VDSR). On the one hand, the layers in the superimposed image are combined in sophisticated degradation ways. Considering the intrinsic relationships between the superimposed image and the degradation patterns, we introduce a variational degradation module to describe the relationship by learning the degradation representations in the latent space. Such additional prior enables the network to handle various degradation processes adaptively. On the other hand, different structural information is stored in the various layers of the superimposed image. To preserve the gradient structure of individual layers, we propose a structure-guided learning strategy that focuses on the structure with superimposed edges by extracting structure-guided knowledge from superimposed images.

To demonstrate the proposed model's effectiveness, we first use five image decomposition methods on two datasets for comparison. Subsequently, we apply our method to a variety of computer vision tasks. Extensive experiments are conducted on three different tasks, including image deraining, image reflection removal, and image shadow removal.

2. Related Work

2.1. Superimposed image decomposition

Superimposed Image decomposition is a general task that covers a wide range of computer vision and computer graphics tasks. An unsupervised method named "Double-DIP" [10] was proposed for image decomposition via coupled "Deep-image-Prior" (DIP) networks [35]. Deep Generative Priors [20] presented a Bayesian approach to image decomposition using a generative model as a prior. However, their methods are limited to handling the input with regular mixed patterns. Deep Adversarial Decomposition (DAD) [51] introduced a novel discriminative network to improve the layer separation. A crossroad L_1 loss was introduced to calculate the loss in a cross-wise manner. G-DPS [41] proposed a dual decoupling network with a perceptual-based training strategy for separation. However, there is still some residual information between the two separated images when using a single-lane structure. BIDeN [14] proposed a multi-lane structure, which separates superimposed images by using one encoder and multiple decoders. Deep-Masking Generative Network [7] presented iterative utilization of residual deep-masking cells to control information propagation, and further a refinement strategy for generation. While showing convincing improvements in many low-level vision tasks, they fail to preserve structural information during complicated degradation.

2.2. Learning degradation representations

Since image restoration involves dealing with various complicated degradation patterns, degradation representations have been developed and employed as an essential component in image denoising and super-resolution. For image denoising, noise variance is commonly used to represent degradation. Many methods [49, 45, 24] were proposed to first estimate the noise variance conditioned on the input noise image. Similarly, for image super-resolution, degeneration representations include Gaussian blur, motion blur, noise, etc. In order to model the blurring process, the model [1, 25] with linear convolutional layers was used, trained with an adversarial loss. Recently, the method [46] elab-

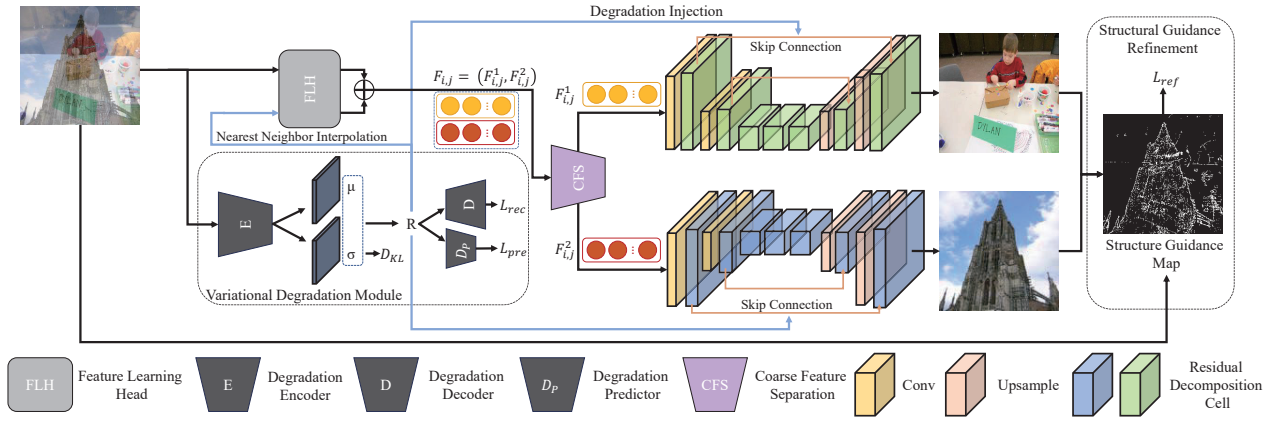


Figure 3. An overview of our method. Our method sequentially performs three stages. Firstly, a variational degradation module is proposed to estimate the degradation representation R, which is constrained by three losses: L_{rec}, L_{pre}, and D_{KL}. Secondly, the feature learning head (FLH) first extracts features from the input image and the predicted degeneration prior. The features are fused together, then roughly divided into two branches through the coarse feature separation (CFS). Then a degeneration injection network is introduced to perform more accurate feature decomposition with the residual decomposition cell (RDC) in the branch. Thirdly, the structural guidance refinement module is responsible for refining the separated images with structure guidance maps.

orately modeled the degradation process based on the perspectives of blur kernels and noise. Although explicit degeneration representations show convincing improvements, they are only task-specific and lack generalization to different tasks. In contrast, our method estimates degeneration representations in the unified framework.

3. Method

3.1. Overview

Given a single superimposed image $I \in [0,1]^{H\times W\times 3}$ containing two image views, we aim to separate it into the corresponding source images x_1 and x_2. Figure 3 shows the overall pipeline of our proposed image separation model - Variational Degenerate to Structural Refinement (VDSR). Our VDSR sequentially performs three stages: (1) estimation of degradation representations. (2) image separation with learned degradation representations. (3) structural guidance refinement.

3.2. Variational Degradation Module

Different degradation processes can result in varying degrees of damage to the texture, color, and contrast of the image [22]. Hence, the superimposed image should have an inherent relationship with the degradation patterns. To address this issue, we propose a variational degradation module (VDM), which is capable of adapting to various complex degradation patterns by fully digging the degradation representations in the latent space. Specifically, we first describe the VDSR using the joint distribution $p(x, I)$ between the source x and superimposed images I. Then we introduce a degeneration representation R which contains prior information about the superposition process. We learn such degeneration representation through a posterior inference $p(R|x, I)$. However, this inference is intractable.

We solve the inference problem based on the variational approximation since the potential distribution of degradation can be significantly controlled and modeled. Accurately, we construct a new encoding distribution, $q(R|I)$, to approximate $p(R|x, I)$. Then, the joint probability distribution $p(x, I)$ can be expressed as follows:

$$\log p(x, I) = \mathcal{L}_b + D_{KL}(q(R|I)\|p(R|x, I)), \quad (2)$$

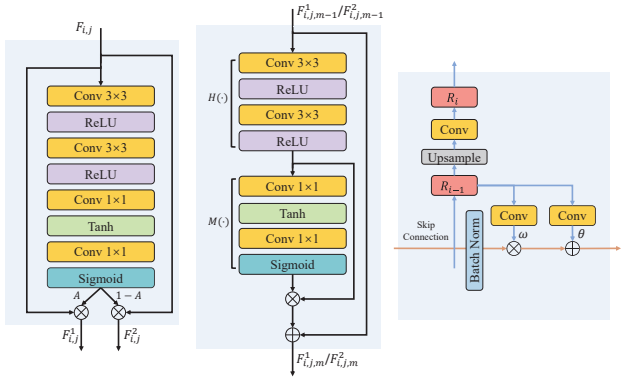
where $\mathcal{L}_b$ represents the variational lower bound and D_{KL} stands for the Kullback-Leibler (KL) divergence.

Due to the non-negative nature of the KL divergence, $\mathcal{L}_b$ constitutes the lower bound of $logp(x, I)$, which is commonly called the evidence lower bound (ELBO). It can be formulated as $\log p(x, I) \geq \mathcal{L}_b$. In this way, by maximizing ELBO, we can naturally approximate the true posterior probability $p(R|x, I)$ by using $q(R|I)$. ELBO is defined as:

$$\mathcal{L}_b = -D_{KL}(q(R|I)\|p(R)) + \mathbb{E}_{R\sim q(R|I)}[\log p(I|R)]. \quad (3)$$

Here $\mathbb{E}_R[\cdot]$ denotes the expectation. We employ three parameterized convolutional neural networks $E(\cdot)$, $D(\cdot)$ and $D_p(\cdot)$ to jointly train the objective function.

The first term is a regularization term that encourages the learned distribution $q(R|I)$ to resemble the true prior distribution $p(R)$. The distribution of $p(R)$ is set to follow a unit Gaussian distribution with zero mean and unit variance.



(a) Coarse Feature Separation (b) Residual Decomposition Cell (c) Degradation Injection Network

Figure 4. The structure of the three submodules.

$$D_{KL}(q(R|I)\|p(R)) = D_{KL}(\mathcal{N}(\mu,\Sigma)\|\mathcal{N}(0,I)) = \frac{1}{2}\left[\sum_{i=1}^{n}\left(-\log\sigma_i^2 - 1 + \sigma_i^2 + \mu_i^2\right)\right], \tag{4}$$

where σ and μ^2 represent the mean and variance, respectively. n refers to the random variable's dimension.

The second term represents the reconstruction likelihood. We first adopt the standard L_1 loss and an adversarial loss [12] as L_{rec}, measuring the distance between the reconstructed output and the superimposed input. Furthermore, for the known degeneration priors, we introduce a prediction loss to constrain the degeneration representation. Specifically, we perform the loss by $D_p(\cdot)$, which consists of two layers of convolution and adaptive max pooling. The prediction loss is described as:

$$L_{pre} = \|Y - D_p(R)\|_1, \tag{5}$$

where Y represents the degradation prior. For example, in the linear superposition , $Y \in \mathbb{R}^{1\times N}$ denotes the mixing factor multiplied in front of the N source images. When we don't have the degeneration prior, like a rainy image, we do not use it.

Overall, the total degradation loss for is the sum of three terms, as follows, we empirically set $\beta_1 = 0.005$.

$$\mathcal{L}_{de} = \beta_1 D_{KL} + \mathcal{L}_{rec} + \mathcal{L}_{pre} \tag{6}$$

3.3. Image Separation with Learned Degradation Representations

In the stage of image separation, we combine the degradation representation R of the previous stage to roughly divide the input features into two branches that produce individual layers for downstream processing.

Formally, given a degradation representation R and an input image polluted by superimposing. We recover R to the input dimensions using nearest neighbor interpolation. Both of them are then projected into the deep embedding feature space through the Feature Learning Head (FLH), which employs a hypercolumn technique to enhance useful features [50]. Subsequently, their fused features $\left(F^1_{i,j}, F^2_{i,j}\right)$ are obtained by summation operation and fed into Coarse Feature Separation (CFS) to be separated into the corresponding constituents of the two branches. To estimate the relevant regions corresponding to various constituents, CFS employs a spatial attention mechanism [19] (see Figure 4(a)). The attention map $\mathbf{A}$ is acquired using the attention's nonlinear function f_{att} and sigmoid function σ.

$$\begin{aligned} \mathbf{A} &= \sigma\left(f_{att}\left(F_{i,j}\right)\right) \\ F^1_{i,j} = F_{i,j} \odot \mathbf{A}, F^2_{i,j} &= F_{i,j} - F^1_{i,j}, \end{aligned} \tag{7}$$

where $F^1_{i,j}$ and $F^2_{i,j}$ denote the separated features of two branches. i and j represent pixel positions.

Furthermore, RDC utilizes the success of the Residual Deep-Masking Cell [7] to progressively reconstruct the images of two layers. It is customized as the core operating unit, sharing structure but not parameters in two branches. The structure of RDC is shown in Figure 4(b). Given an input feature gained from the $(m-1)$-th RDC, the feature is first refined by $H(\cdot)$ containing convolutional layers and ReLU. Then the mask with the interval [0,1] is learned by $M(\cdot)$. In addition, a residual connection is used to avoid the vanishing gradient problem. The RDC is expressed as:

$$F^1_{i,j,m} = F^1_{i,j,m-1} + M(H(F^1_{i,j,m-1})) \odot H(F^1_{i,j,m-1}), \tag{8}$$

where $\odot$ denotes the element-wise multiplication.

Injecting the degradation representations. Our RDC introduces the degradation injection network (DInet) in order to effectively incorporate the degradation representation R to perform layer separation. Specifically, the orange arrow represents forward propagation in skip connections and the blue arrow shows degenerate forward propagation in Figure 4(c). R_i is obtained by performing upsampling and convolution on R_{i-1}. DInet conducts affine transformation by scaling and shifting feature with ω and θ obtained by convolution.

$$F_n = \omega_n \odot f_n + \theta_n, \tag{9}$$

where F_n is the modulated skip-connection Features f_n in the n-th skip-connection.

3.4. Structural Guidance Refinement

To preserve the gradient structure and refine the separated images, we propose a structure-guided learning strategy, which assists the network in focusing on the geometric structure by learning additional structure guidance maps.

Concretely, the structure guide map is extracted from the input image I, which is capable of detecting the pixel where texture-copy or artifacts of superposition occurs. Based on the assumptions that these pixels are sparse, we use the L_0 norm rather than the L_1 norm to penalize the number of non-zero vectors in the structure guide map. The L_0 norm has been shown to be a good choice for image structure in many methods [3, 2, 27]. By introducing a binary gate $\mathcal{G}$, the L_0 regularization of the structure guidance map $\mathcal{M}$ can be expressed as:

$$\|\mathcal{M}\|_0 = \sum_l \mathcal{G}_l, \tag{10}$$

where l represents a pixel. $\mathcal{G}_l$ determines whether texture-copy exists at pixel l ($\mathcal{G}_l$ = 1 means there exists). However, the optimization is computationally intractable due to the non-differentiability of the L_0 norm.

Hard concrete distribution. Since L_0 is difficult to converge, we use a hard concrete distribution to relax the discrete nature [28], which allows gradients to flow efficiently in L_0. Specifically, it is defined as $\mathcal{H}(s|\phi)$ with parameters $\phi = (\ell, \tau)$, where s is a random variable distributed between 0 and 1. The probability of the gate opening ($\mathcal{G} = 1$) is determined by ℓ. τ determines the approximate distribution of $\mathcal{H}(s|\phi)$. We empirically set τ = 0.5. After stretching such distribution into the interval (γ, ζ) with $\gamma = -0.1$ and $\zeta = 1.1$, we apply the hard-sigmoid on its random samples.

$$\begin{aligned} &u \sim \mathcal{U}(0,1), s = \text{Sigmoid}((\log u - \log(1-u) + \ell)/\tau) \\ &\bar{s} = s(\zeta - \gamma) + \gamma, \mathcal{G} = \min(1, \max(0, \bar{s})). \end{aligned} \tag{11}$$

In this way, the original binary gate is represented as a soft gate $\mathcal{G} \in [0, 1]$. It provides a better approximation of discrete properties and allows gradient-based optimization.

Structure guidance map. To get the map $\mathcal{M}$, we propose the following losses to optimize, including sparse term $\mathcal{L}_{spa}$ and structure term $\mathcal{L}_{str}$, where we set $\beta_2 = 0.2$.

$$\begin{aligned} \mathcal{L}_{\mathcal{M}} =& \beta_2 \mathcal{L}_{spa} + \mathcal{L}_{str} \\ \mathcal{L}_{spa} =& \sum_l \frac{1}{\|\nabla I_l\|} p\left(\mathcal{G}_l \neq 0\right), \end{aligned} \tag{12}$$

where ∇ denotes the gradient operator. $p\left(\mathcal{G}_l \neq 0\right)$ represents the likelihood that the gate at pixel l is non-zero.

The structure term is used to avoid all gates becoming 0. In general, pixels with large gradients are assigned higher since they are more likely to be caused by texture copy, while pixels with small gradients are assigned lower since they are more likely to arise from the same layer. The structure term is defined as follows.

$$\mathcal{L}_{str} = \sum_l \left(\|\nabla I_l\| \left(1 - \mathcal{G}_l\right) + \frac{1}{\|\nabla I_l\|} \mathcal{G}_l \right) \tag{13}$$

According to Eq. 11, we can calculate the value of the gate $\mathcal{G}$ and obtain the structure guidance map $\mathcal{M}$.

3.5. Overall Objective Function

The objective function of the VDSR contains four terms: a degradation loss, a pixel loss, an adversarial loss, and a refinement loss.

Degradation loss. To ensure that the degradation representation R can learn the degradation information from the superimposed images, the loss is applied to constrain R.

Pixel loss. Instead of implementing L_1 loss in the cross-road way [51], we present a fixed order of two outputs (both two orders are feasible), then calculate the distance with their corresponding ground-truths, where $\hat{x}_i$ denotes output and $i \in \{1, 2\}$.

$$\mathcal{L}_{\text{pixel}} = \mathbb{E}_{(\hat{x}_i, x_i) \sim p_i(\hat{x}_i, x_i)} \left\{ L_1\left(\hat{x}_1, x_1\right) + L_1\left(\hat{x}_2, x_2\right) \right\} \tag{14}$$

Conditional Adversarial loss. The conditional adversarial loss assists models in achieving high separation performance. We adopt a similar structure of discriminator D as Pix2Pix [18].

$$\begin{aligned} \mathcal{L}_{adv} =& -\mathbb{E}_{(x_i, I) \sim p_i(x_i, I)} \left\{ \log\left(D\left(\hat{x}_i, I\right)\right) \right\} \\ \mathcal{L}_D =& \mathbb{E}_{(\hat{x}_i, I) \sim p_i(\hat{x}_i, I)} \left\{ \log D\left(\hat{x}_i, I\right) \right\} - \\ & \mathbb{E}_{(x_i, I) \sim p_i(x_i, I)} \left\{ \log D\left(x_i, I\right) \right\} \end{aligned} \tag{15}$$

Refinement loss. we propose a novel loss to preserve geometric structure information under the guidance of structure guide map $\mathcal{M}$.

$$\mathcal{L}_{ref} = \|\mathcal{M} \odot (\nabla x_i - \nabla \hat{x}_i)\|_2^2 \tag{16}$$

To sum up, our final loss function is formulated as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{de} + \lambda_2 \mathcal{L}_{pixel} + \lambda_3 \mathcal{L}_{adv} + \lambda_4 \mathcal{L}_{ref}, \tag{17}$$

where the coefficients are experientially set as $\lambda_3 = 0.01$, $\lambda_1 = \lambda_2 = \lambda_4 = 1$.

3.6. Training details

Our VDSR is implemented by PyTorch with an NVIDIA RTX 3080Ti. We use ADAM as the optimizers with $\beta_1 = 0.5$, and $\beta_2 = 0.999$, and the initial learning rate is set to 2×10^{-4}. The learning rate is adjusted by the linear decay strategy. We train our model for 200 epochs with a batch size of 2. Our training process consists of two optimization stages. First, we obtain the map $\mathcal{M}$ by minimizing $\mathcal{L}_{\mathcal{M}}$. The variable to be optimized is only the ℓ in Eq. 11. Then we use the total loss $\mathcal{L}$ to optimize the two prediction. ∇I in Eq. 12 is obtained following $\nabla I = max\{\nabla I, \epsilon\}$, the ϵ is set to 0.01 in the implementation. The code will be released at https://github.com/lwyfish/Variational-Degeneration-to-Structural-Refinement.

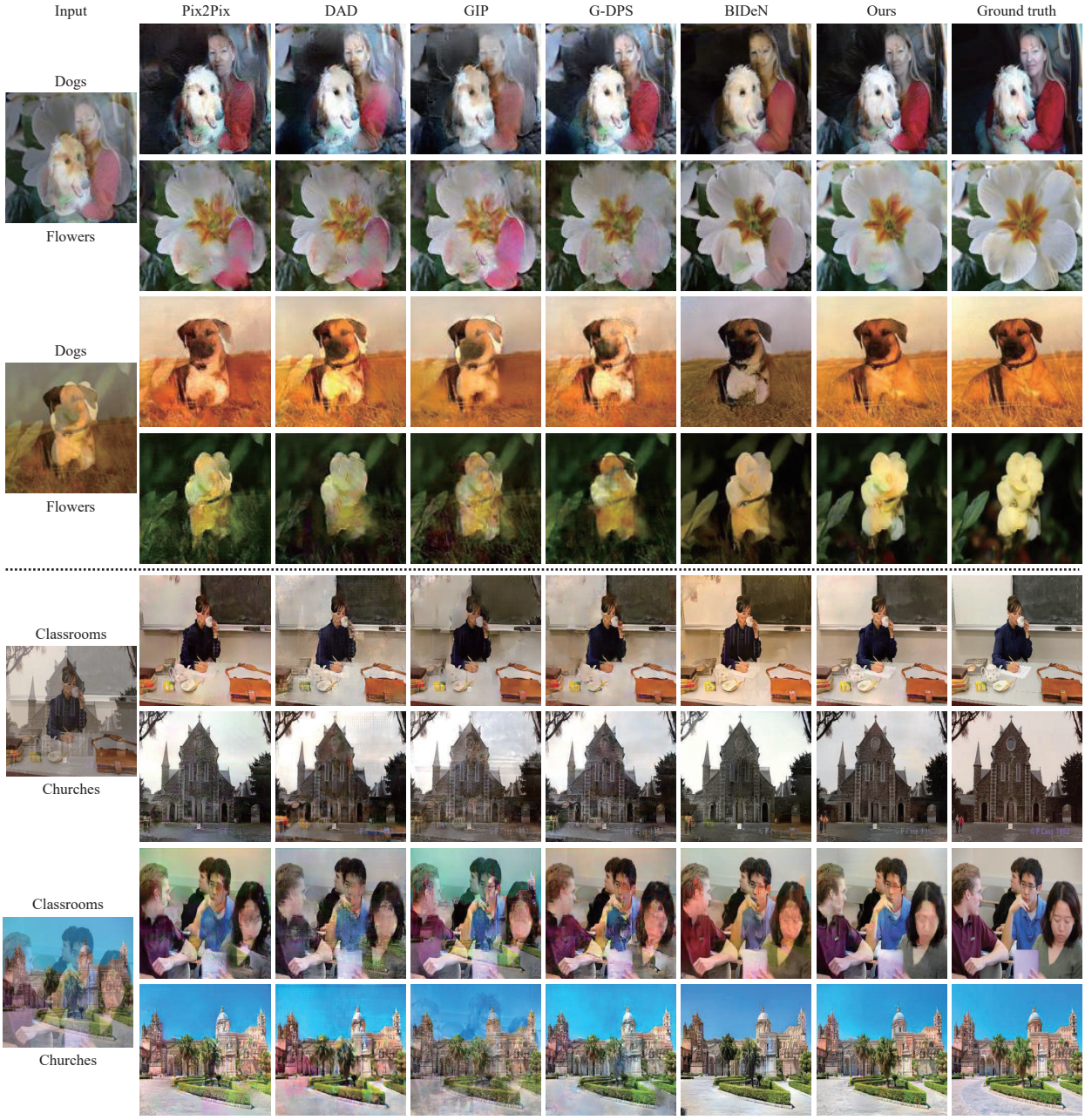


Figure 5. Comparisons with the superimposed image decomposition methods on two mixing datasets. Above the dotted line is the Dogs + Flwrs dataset, and below the dotted line is the LSUN mixed dataset.

4. Experiments

We evaluate the proposed method on eight datasets for four tasks, including 1) superimposed image decomposition, 2) single image deraining, 3) single image reflection removal, and 4) single image shadow removal.

4.1. Superimposed Image Decomposition

Datasets: In our basic task, we use the following two datasets for evaluations, i.e., 1) Stanford-Dogs (Dogs) [23] + VGG-Flowers (Flwrs) [29], 2) LSUN Classroom + LSUN Church [44]. To be specific, we use the original training/testing split in these datasets [23, 29, 44] and follow the same settings as in [51]. During training, we randomly

(a) Input (b) JORDER (c) PreNet (d) VDSR (Ours) (e) Ground truth (f) Rain streak (Ours)

Figure 6. Qualitative comparison of different methods for single image deraining on Rain100H dataset.

Table 1. Performance (PSNR / SSIM) of different methods for superimposed image decomposition on two mixing datasets

Methods	Dogs+Flwrs	Lsun
Pix2Pix [18]	24.04 / 0.703	23.89 / 0.755
DAD [51]	22.83 / 0.669	23.27 / 0.720
GIP [52]	23.28 / 0.6784	22.98 / 0.715
G-DPS [41]	23.11 / 0.685	23.95 / 0.759
BIDeN [14]	24.82 / 0.704	23.50 / 0.705
VDSR (Ours)	**28.15 / 0.851**	**27.98 / 0.869**

Table 2. Ablation studies for each component of VDSR on two mixing datasets. The performance is formatted as PSNR / SSIM.

Ablation	Dogs+Flwrs	Lsun
I	26.78 / 0.821	26.74 / 0.842
II	27.01 / 0.822	26.92 / 0.855
III	27.49 / 0.844	27.95 / 0.872
IV	27.60 / 0.838	27.71 / 0.861
V	27.87 / 0.842	27.82 / 0.863
VI	27.54 / 0.836	27.47 / 0.853
Ours	**28.15 / 0.851**	**27.98 / 0.869**

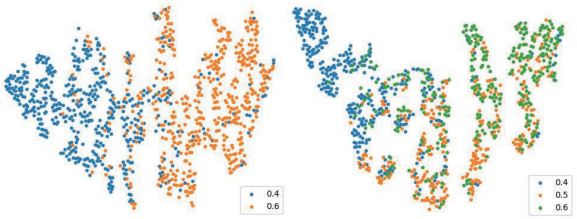
select two images (x_1, x_2) from two subsets of one group and then linearly mix them as $I = \alpha x_1 + (1 - \alpha) x_2$ with a random linear mixing factor α, where $\alpha \in [0.4, 0.6]$. During evaluation, we mix two images with a constant α value of 0.5. In summary, we randomly generate 6000 pairs of Dogs+Flwrs and 6000 pairs of LSUN Classroom+Church at each epoch for training, and fixedly produce 1000 pairs of Dogs+Flwrs and 300 pairs of LSUN Classroom+Church for testing. **Baselines:** We compare the proposed model with other five popular methods for single superimposed image decomposition, including one well-known image translation methods Pixel-to-Pixel [18] and four recent methods DAD [51], GAN for Inverse Problems (GIP) [52], G-DPS [41], BIDeN [14]. Two popular metrics are used for quantitative comparisons [5, 9, 34], namely the Peak Signal-to-Noise Ratio (PSNR) [17] and Structure Similarity (SSIM) [39]. Higher scores indicate better.

Results: Table 1 reports the quantitative comparison results of the five superimposed image decomposition methods in terms of PSNR and SSIM on two mixing datasets. The results demonstrate that our model performs optimally in terms of metrics on both datasets. Besides the dominance in quantitative evaluations, VDSR also shows superiority in qualitative comparisons as shown in Figure 5. It can be observed that our method generates results with rich and credible image textures while unmixing the superimposed images. For other comparison methods, they tend to blur the image content or still leave some artifacts from another layer. For instance, only our VDSR recovers believable image details in the "dog" image, while competing methods fail to perform a clean separation. Although BIDeN can separate well, it still has obvious color distortion.

4.2. Ablation Study

We perform ablation experiments to analyze the effectiveness of each component inside VDSR using five variants. (I) Base. Our base model, which only contains the image separation stage with two separation branches, and the loss functions are $\mathcal{L}_{pixel}$ and $\mathcal{L}_{adv}$. (II) Base + standard gradient loss. In order to evaluate the importance of the structure guidance map, we replace L_{ref} with the standard gradient loss. (III) Base + L_{ref}. We increase the stage of structural guidance refinement, using structure guidance maps. (IV) Base + VDM. The variational degradation module is added on the base model I without using DInet to inject the degradation representation. (V) Base + VDM + DInet. To verify the importance of DInet, we increase the

Figure 7. Visualization of structure guidance map (2nd row). Refined maps provide better structural guidance information.



(c) Two mixing factors (b) Three mixing factors

Figure 8. T-SNE visualizations of degeneration representations.

Table 3. Quantitative results (PSNR / SSIM) of image deraining on the Rain100H [43].

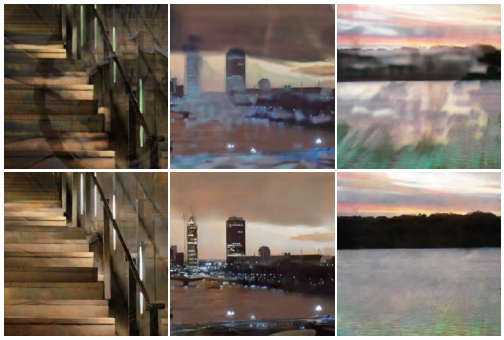
Methods	Rain100H [43]
DDN [8]	22.26 / 0.693
JORDER [43]	23.45 / 0.749
RESCAN [26]	26.45 / 0.846
DID-MDN [48]	25.00 / 0.754
DAF-Net [15]	28.44 / 0.874
PreNet [32]	29.46 / 0.899
BIDeN [14]	29.65 / 0.876
Restormer [47]	**31.46** / 0.904
VDSR (Ours)	30.89 / **0.905**

injection process based on IV. (VI) Base + VDM + L_{ref}. The utilization of DInet alone is unsuccessful due to the lack of degradation representation R. Evaluation is performed on the basic task, i.e., superimposed image decomposition.

Table 2 shows the results of ablation experiments for VDSR. All the ablation studies demonstrate the effectiveness of our proposed structure guidance map and the degradation representation. We visualize the structure guidance map in Figure 7. The map preserves the structural information of the mixing between layers. We also show T-SNE [36] visualization of degradation representation R based on the dataset 4.1 in Figure 8. As we can see, points with similar mixing factors are compact in the latent representation space, and the embedded points vary continuously from low to high factors. It can be proved that R contains degeneration prior, which is suitable for image decomposition.

4.3. Application: Single Image Deraining

For image deraining, we conduct experiments on a diffcult dataset, Rain100H (heavy-rain) [43], which consists of 1254 images for training and 100 images for testing. We compare our model with state-of-the-art methods for single image deraining including JORDER [43], PreNet [32], and Restormer [47], etc. The training and testing of all of these methods follow the same training and testing split outlined above. From Table 3 and Figure 6, one could observe that VDSR achieves the best result in almost all tests. Rostormer shows superior results in PSNR. In contrast, VDSR is better in SSIM due to the structural guidance refinement. Our approach yields promising results in removing rain streaks, enhancing the visibility and preserving details. An additional benefit of our approach is that it enables estimation of the rainstreak map.



(a) Focused (b) Defocused (c) Ghosting

Figure 9. Qualitative results of our method for image reflection removal on the dataset [40]. 1st row: input. 2nd row: output.

4.4. Application: Single Image Reflection Removal

We perform the experiments on three popular benchmark datasets [40]. The dataset contains three types of reflections, including "focus", "defocus" and "ghosting", and each reflection type includes 4000 training images and 100 testing images. Since the ground truths for blurred reflection images are usually not available, we simply set the ground truth for the second output to "zero image" [51] and train only one branch of the VDSR during training. We compare our model with the following state-of-the-art methods for image reflection removal in Table 4, including RmNet [40], BIDeN [14], etc. Figure 9 shows some results which demonstrate that our method can recognize the reflection artifacts and remove them.

4.5. Application: Single Image Shadow Removal

In this experiment, we validate the performance of the proposed method on two commonly used datasets: SRD [31] and ISTD [38]. These two datasets are composed of 3088 and 1870 shadow/shadow-free image pairs captured

Table 4. Quantitative results (PSNR / SSIM) of image reflection removal on the dataset [40].

Methods	Focused set	Defocused set	Ghosting set
CEILNet [6]	19.524 / 0.742	20.122 / 0.735	19.685 / 0.753
Zhang *et al.* [50]	17.090 / 0.712	18.108 / 0.758	17.882 / 0.738
BDN [42]	14.258 / 0.632	14.053 / 0.639	14.786 / 0.660
RmNet [40]	21.064 / 0.770	22.896 / 0.840	21.008 / 0.780
BIDeN [14]	23.007 / 0.801	23.707 / 0.813	23.866 / 0.817
VDSR (Ours)	**26.226 / 0.915**	**26.549 / 0.933**	**26.579 / 0.924**

Table 5. Quantitative results of image shadow removal on SRD [31] and ISTD [38], using root mean square error (RMSE) (lower is better) [13].

Methods	SRD [31]	ISTD [38]
Gong *et al.* [11]	8.730	8.530
DeshadowNet [31]	6.640	7.830
ST-CGAN [38]	8.230	7.470
DSC [16]	6.210	6.670
ARGAN [4]	5.740	6.680
BIDeN [14]	6.610	7.750
SG-ShadowNet [37]	**4.230**	**3.40**
VDSR (Ours)	4.984	6.063

in real environments, respectively. In Table 5, we compare our proposed method with some task-specific competitive methods, including DSC [16] and SG-ShadowNet [37], etc. Constrained by unified settings and the absence of priors for shadow removal, VDSR does not show superior quantitative results compared to SG-ShadowNet designed for shadow removal tasks only. Figure 10 shows the qualitative comparison of our methods with DSC+ on the above two datasets. One can see that our proposed method can eliminate the shadow without obvious artifacts.

(a) Input (b) DSC+ (c) Ours (d) Ground truth

Figure 10. Qualitative comparison for image shadow removal on the SRD [31] (1st row) and ISTD [38] (2nd row).

5. Conclusion

In this paper, a novel unified framework called VDSR is proposed for single superimposed image decomposition. Using degeneration representations in the latent space, we propose a variational degradation module to mine the intrinsic relationship between superimposed images and degradation patterns. This additional information can help adaptively handle various degradation patterns. Additionally, to solve the problem of blurred edges caused by structural distortions, we propose the structure guidance map to assist the model in learning an informative structural representation. Quantitative and qualitative results on four tasks have shown the effectiveness of our proposed method.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 62101378 and Grant 62171318, National Key Research and Development Program of China under Grant 2021YFE0204200, Tianjin Transportation Science and Technology Development Plan (Project No. 202234), and China Postdoctoral Science Foundation under Grant 2023M732612.

References

[1] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Rick Chartrand and Valentina Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24(3):035020, 2008.

[3] Huasong Chen, Zhenhua Xu, Qiansheng Feng, Yuanyuan Fan, and Zhenhua Li. An l0 regularized cartoon-texture decomposition model for restoring images corrupted by blur and impulse noise. *Signal Processing: Image Communication*, 82:115762, 2020.

[4] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10213–10222, 2019.

[5] Yu Dong, Yihao Liu, He Zhang, Shifeng Chen, and Yu Qiao. Fd-gan: Generative adversarial networks with fusion-discriminator for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10729–10736, 2020.

[6] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3238–3247, 2017.

[7] 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*, 30:4867–4882, 2021.

[8] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3855–3863, 2017.

[9] Xueyang Fu, Borong Liang, Yue Huang, Xinghao Ding, and John Paisley. Lightweight pyramid networks for image deraining. *IEEE transactions on neural networks and learning systems*, 31(6):1794–1807, 2019.

[10] Yosef Gandelsman, Assaf Shocher, and Michal Irani. " double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11026–11035, 2019.

[11] Han Gong and Darren Cosker. Interactive shadow removal and ground truth for variable scene categories. In *BMVC*, pages 1–11, 2014.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[13] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2956–2967, 2012.

[14] Junlin Han, Weihao Li, Pengfei Fang, Chunyi Sun, Jie Hong, Mohammad Ali Armin, Lars Petersson, and Hongdong Li. Blind image decomposition. In *European Conference on Computer Vision*, pages 218–237. Springer, 2022.

[15] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2019.

[16] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2795–2808, 2019.

[17] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[20] Vivek Jayaram and John Thickstun. Source separation with deep generative priors. In *International Conference on Machine Learning*, pages 4724–4735. PMLR, 2020.

[21] Seo-Won Ji, Jeongmin Lee, Seung-Wook Kim, Jun-Pyo Hong, Seung-Jin Baek, Seung-Won Jung, and Sung-Jea Ko. Xydeblur: Divide and conquer for single image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17421–17430, 2022.

[22] Kui Jiang, Zhongyuan Wang, Zheng Wang, Chen Chen, Peng Yi, Tao Lu, and Chia-Wen Lin. Degrade is upgrade: Learning degradation for low-light image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1078–1086, 2022.

[23] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2. Citeseer, 2011.

[24] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 736–753. Springer, 2022.

[25] Shang Li, Guixuan Zhang, Zhengxiong Luo, Jie Liu, Zhi Zeng, and Shuwu Zhang. From general to specific: Online updating for blind super-resolution. *Pattern Recognition*, 127:108613, 2022.

[26] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 254–269, 2018.

[27] Zhetong Liang, Jun Xu, David Zhang, Zisheng Cao, and Lei Zhang. A hybrid l1-l0 layer decomposition model for tone mapping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4758–4766, 2018.

[28] Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

[29] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006.

[30] Surya N Patnaik, James D Guptill, and Dale A Hopkins. Subproblem optimization with regression and neural network approximators. *Computer methods in applied mechanics and engineering*, 194(30-33):3359–3373, 2005.

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

[32] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3937–3946, 2019.

[33] 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*, pages 234–241. Springer, 2015.

[34] Chunwei Tian, Yong Xu, and Wangmeng Zuo. Image denoising using deep cnn with batch renormalization. *Neural Networks*, 121:461–473, 2020.

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[37] Jin Wan, Hui Yin, Zhenyao Wu, Xinyi Wu, Yanting Liu, and Song Wang. Style-guided shadow removal. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 361–378. Springer, 2022.

[38] Jifeng Wang, Xiang Li, and Jian 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*, pages 1788–1797, 2018.

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[40] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3771–3779, 2019.

[41] Yan Xu, Wenyu Li, Yang Yang, Haoran Ji, Beichen Li, and Yue Lang. Multiple targets echo separation on radar range-doppler maps via dual decoupling perception. *IEEE Sensors Journal*, 2022.

[42] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proceedings of the european conference on computer vision (ECCV)*, pages 654–669, 2018.

[43] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1357–1366, 2017.

[44] 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.

[45] Zongsheng Yue, Hongwei Yong, Qian Zhao, Deyu Meng, and Lei Zhang. Variational denoising network: Toward blind noise modeling and removal. *Advances in neural information processing systems*, 32, 2019.

[46] Zongsheng Yue, Qian Zhao, Jianwen Xie, Lei Zhang, Deyu Meng, and Kwan-Yee K Wong. Blind image super-resolution with elaborate degradation modeling on noise and kernel. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2128–2138, 2022.

[47] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[48] He Zhang and Vishal 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*, pages 695–704, 2018.

[49] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.

[50] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4786–4794, 2018.

[51] Zhengxia Zou, Sen Lei, Tianyang Shi, Zhenwei Shi, and Jieping Ye. Deep adversarial decomposition: A unified framework for separating superimposed images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12806–12816, 2020.

[52] Zhengxia Zou, Tianyang Shi, Zhenwei Shi, and Jieping Ye. Adversarial training for solving inverse problems in image processing. *IEEE Transactions on Image Processing*, 30:2513–2525, 2021.