Deep Semantic Statistics Matching (D2SM) Denoising Network

Kangfu Mei, Vishal M. Patel, and Rui Huang

[1] Johns Hopkins University
[2] The Chinese University of Hong Kong, Shenzhen
https://kfmei.page/d2sm

Abstract. The ultimate aim of image restoration like denoising is to find an exact correlation between the noisy and clear image domains. But the optimization of end-to-end denoising learning like pixel-wise losses is performed in a sample-to-sample manner, which ignores the intrinsic correlation of images, especially semantics. In this paper, we introduce the Deep Semantic Statistics Matching (D2SM) Denoising Network. It exploits semantic features of pretrained classification networks, then it implicitly matches the probabilistic distribution of clear images at the semantic feature space. By learning to preserve the semantic distribution of denoised images, we empirically find our method significantly improves the denoising capabilities of networks, and the denoised results can be better understood by high-level vision tasks. Comprehensive experiments conducted on the noisy Cityscapes dataset demonstrate the superiority of our method on both the denoising performance and semantic segmentation accuracy. Moreover, the performance improvement observed on our extended tasks including super-resolution and dehazing experiments shows its potentiality as a new general plug-and-play component.

Keywords: Denoising, Semantic Segmentation, Super-resolution, Dehazing, Implicit Modeling, Score Matching

1 Introduction

Deep learning based methods [9, 24, 62] have achieved a dramatic leap in the performance of various image restoration tasks. Typically, they employ a Convolutional Neural Network (CNN) on a set of image pairs, consisting of degraded images and corresponding clear images, for restoration learning. By maximizing the correspondence between each pair of the CNN-restored results and the clear image, the CNN is trained to map images from the degraded domain into the clear domain. However, blur issues always existed in such a manner. Recent work [17] called perceptual loss finds that maximizing the correspondence in the semantic feature space of pre-trained large-scale classification network (*e.g.*, VGG [50] network trained on ImageNet [8]) leads to better visual quality [35,63]. A more widely used strategy inspired by GANs [12], which employs a discriminator to implicitly enforce the distribution of restored images to be consistent with
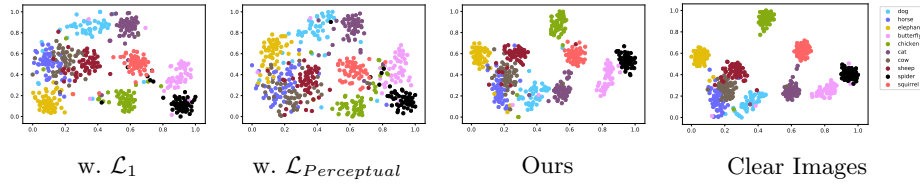
Fig. 1: **t-SNE of Denoised Images in the Semantic Feature Space** By exploiting t-SNE [34] to reduce dimensions of semantic features and project them into 2D coordinates, we visualize the distributions of denoised animal images in the semantic feature space. Ours preserves most semantics as the clear images.

the distribution of clear images in terms of KL- and JS- divergence, can largely improve the perceptual quality of restored images. But the training procedure is often unstable, mostly because the objective is a zero-sum non-cooperative game that cannot be easily solved. Thus, it is straightforward to wonder whether it is possible to combine the pre-trained large-scale networks in an adversarial or statistical manner to bypass their drawbacks and avail their advantages together?

To answer the above question, we first look at the training procedure from a probabilistic view in the semantic feature space (the space of extracted semantic features of each images), which contains many clusters of different semantics. Figure 1 visualizes these clusters with t-SNE [34] in animals images selected from ImageNet [8]. An image that belongs to a specific cluster and owns intrinsic semantics is called Single-Semantic Image in this paper. For example, animal images are single-semantic images, because they have common semantics of species, even though individual animals with the same specie look different. Intuitively, restored single-semantic images should preserve the same distribution as the corresponding clear single-semantic images in the semantic feature space. However, the objectives of existed methods for denoising learning cannot preserve this, as w.$\mathcal{L}_1$ and w.$\mathcal{L}_{Perceptual}$ show in Figure 1. Therefore, we argue that minimizing the divergence between the probability distributions estimated in the restored domain and the one estimated in the clear domain should be a potential promising solution. Similar idea is also validated by MMDGAN [22] and its variants for generating face and bedroom images, but the idea is rarely researched in the restoration literature.

Different from the single-semantic image, natural images like a cityscape image often consists of multiple objects, and can not be easily identified by a single semantic, here it is called Complex-Semantic image. The semantic feature extracted from such an image may not belong to any simple semantic clusters, but resides on an extremely complicated manifold in the semantic space. For example, cityscape images usually consist of objects/regions of different semantics like *road*, *sidewalk*, and *building*, as classified by the Cityscapes dataset [7]. To approximate and compare the probability distributions of complex-semantic images is nontrivial due to their unique intrinsic uncertainty of semantics.

In the paper, we propose a new distribution-wise objective for denoising learning, and is capable of being extended into general restoration tasks, towards both the single-semantic images and complex-semantic images. It learns to preserve the probability distribution of denoised images in the semantic feature space, and it is called as Deep Semantic Statistics Matching (D2SM) Denoising Network. The objective of D2SM exploits a way similar to Kernel Density Estimation (KDE) [47] to implicitly estimate the probability distributions of semantic features from a set of denoised images and clear images, and then Kullback-Leibler (KL) divergence between two distributions is used as the objective. Here, one of our major novelty comes from the way of density estimation, where we model the probability distributions based on internal patches from a single complex-semantic image or multiple single-semantic images. The way of availing internal patches tends to be more appropriate than modeling the distribution of multiple complex-semantic images. Such a phenomenon is also proved in recent work [66] suggests that the internal visual entropy of a single image is much smaller than multiple images. Therefore, we propose to use the divergence of patch distributions to guide the learning, called Patch-Wise Internal Probability.

Nevertheless, statistically estimating the density of patches conducted in a single mini-batch usually requires a great large number of samples. To maintain the trade-off between the computational cost and accuracy, another major novelty of the paper, called Memorized Historic Sampling, is proposed, inspired by recent contrastive learning related works [14, 60]. By simply leveraging the statistics among the mini-batch and memorized historic mini-batch in queues, we demonstrate that D2SM significantly outperforms the same network backbone with perceptual loss and other state-of-the-art objectives, without additional information or parameters. Empirical evaluation validates that D2SM largely improves not only the effectiveness of denoising, but also super-resolution and dehazing, and hence it should be able to be generally applied to different tasks and network architectures.

Our contributions are therefore three-fold:
(i) We propose D2SM for image denoising learning, which minimizes the distribution divergence instead of the sample-to-sample distance in the semantic feature space. (ii) D2SM is adapted to complex-semantic images in a patch-wise manner, which can decompose complex semantics in natural images for efficient distribution approximation. (iii) Extensive experiments are conducted to demonstrate that D2SM substantially outperforms the original perceptual loss and other state-of-the-art losses, without modifying the network architecture or accessing the additional data. The superior accuracy in high-level vision tasks further validates that D2SM indeed transfers semantics for restoration.

2   Related Work

Resulted by the emergence of deep neural networks, recent CNN based methods have led to a dramatic leap in image restoration. Among them, most works

utilize the pixel-wise similarity metrics as their objective, *e.g.*, $\mathcal{L}_1$ and $\mathcal{L}_{MSE}$. Though higher performance in metrics like PSNR or SSIM [59] is achieved by using these loss functions, recent work [63] finds that these metrics do not reflect human perceptual preferences. In contrast, results generated by CNNs trained with the perceptual objective are more closely correlated with the human judgment [35]. These methods measure the similarity of two images in the pre-trained high-level vision networks, usually VGG classification network [50] trained in ImageNet [8]. Different perceptual objectives have been proposed in this category, *e.g.*, $\mathcal{L}_{MSE}$ of features [11, 17, 20, 38, 58, 63], contextual objective [35, 36], and semantic label [27, 28]. However, these methods lack a reasonable explanation of the effectiveness led by the perceptual objective [63]. Furthermore, the frozen network pre-trained on certain datasets, *e.g.*, ImageNet, is not appropriate for the image restoration tasks conducted on the large-scale, diverse natural image datasets [1] or specific semantic image datasets [7, 19, 32]. Here we hypothesize that these issues come from the objectives that estimate the sample-to-sample distance in the feature space. By exploiting the characteristics of single-semantic patches from natural images, which can be associated with an embedded manifold, we implicitly measure the divergence of the probability distributions estimated from restored images and clear images in the semantic feature space, and we use it as the objective to bypass the above issues.

Similar ideas that minimize the distribution divergence instead of the sample-to-sample distance have been proposed before. In the area of image restoration, Contextual loss [35, 36] that proposed for misaligned image transformation implicitly minimizes the divergence between restored images and clear images. It approximates the divergence by the contextual relationships within patches from a single image (*i.e.* single image statistics [66]), and hence enables image-to-image translation to be conducted on the misaligned image pairs. However, its performance is usually limited by the low accuracy of feature matching [64] and leads to worse restoration performance in aligned image restoration learning. More similar works that avail statistical features are GMMN [26] and GFMN [46]. They achieve the generative ability without adversarial learning in the problematic min/max game. Nevertheless, they are not designed for the image restoration learning that majorly consists of natural images with diverse appearance, and the desired superiority cannot be gained here. In the area of domain adaption, minimizing the statistics feature difference of the high-level vision networks can help networks adapt to unseen domain directly, like CORAL [51] and MMD [53]. However, these methods require semantic labels, which is not practical in the real-world image restoration datasets. By exploiting the internal statistics [66] of natural images, our proposed method successfully facilitates the restoration learning through more accurate divergence approximation.

3   Method

Let $\mathcal{X} \subset \mathbb{R}^{H \times W \times C}$ denotes the domain of degraded images caused by factors like noising, and $\mathcal{Y} \subset \mathbb{R}^{H \times W \times C}$ denotes the domain of corresponding clear images.
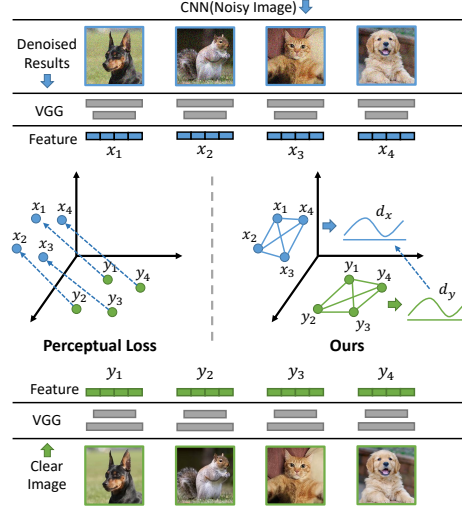
Fig. 2: **Perceptual loss vs. Ours.** We minimize the distribution divergence between a set of restored images and the corresponding clear images, instead of the sample-to-sample distance, in the semantic feature space (*e.g.* the penultimate layer of VGG). This procedure better simplifies the restoration learning and ameliorates underfitting compared with the perceptual loss.

We wish to restore $x \in \mathcal{X}$ to appear like its corresponding target image $y \in \mathcal{Y}$ by using a denoising network $G(\cdot)$ that outputs $\tilde{y} = G(x)$. To force the outputs $\tilde{y}$ maintains as much the perceptual detail as possible, recent works [11,17,20,35,36] exploit the pre-trained high-level vision networks (*e.g.*, the intermediate layer of VGG), denoted as $\Phi(\cdot)$, to guide the restoration learning by minimizing the similarity between $\tilde{y}$ and y in the feature space of $\Phi(*)$. This can be formulated as the objective with the similarity metric $D(\cdot)$:

$$\mathcal{L}(x, y, G) = D(\Phi(y), \Phi(G(x))). \tag{1}$$

In practical, the similarity metric $D(\cdot)$ is usually implemented by Mean Square Error (MSE) or Contextual Distance [36].

Contrastively, we take the denoising learning as minimizing the divergence of probability distributions estimated by denoised images and clear images in the semantic feature space. Given N samples of image pairs that consist of $T_x = \{x_1, x_2, \ldots, x_N\}$ and $T_y = \{y_1, y_2, \ldots, y_N\}$, we incorporate the mutual information [54] of them in the feature space of $\Phi(G(\cdot))$ into the restoration learning. Such a manner is empirically proven to be effective to facilitate knowledge transferring [5,31,40–43,55]. By minimizing the divergence of the estimated probability distribution between samples T_x in $\Phi(G(\cdot))$ and T_y in $\Phi(\cdot)$, denoted as $\mathcal{G}'$ and $\mathcal{G}$, we force $G(\cdot)$ to better maintain the geometry of the feature space $\Phi(\cdot)$ estimated in the clear image domain $\mathcal{Y}$. In doing so, we formulate the final

objective as

$$\mathcal{L}(T_x, T_y, G) = \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} g'_{j|i} \log(\frac{g'_{j|i}}{g_{j|i}}). \tag{2}$$

To elaborate, we will detail the divergence approximation in Section 3.1, and the sampling strategy in Section 3.2 and Section 3.3.

3.1   Probability Distribution Divergence

Here we model the correlation of samples from the same domain in the semantic feature space as the probability distribution. Several methods have been proposed for modeling the correlation, including, but not limited to, probabilistic based [41, 42], embedding based [5, 43], graph based [31], and more [40]. In this work, we exploit the kernel density estimation to estimate the probability distribution of samples in the semantic feature space, which describes the probability of each sample to select its neighbors [34]. It is empirically proven to be effective for describing the geometry of feature space by Passalis et al. [41, 42]. To elaborate, we denote the probability distribution between any two samples i, j from the clear domain as $g_{i|j}$ and the restored domain as $g'_{i|j}$. Based on the extracted feature f^x and f^y from $\Phi(G(\cdot)))$ and $\Phi(\cdot)$, the probability distributions are estimated by:

$$g'_{i|j} = \frac{K_{cosine}(f^x_i, f^x_j)}{\sum_{k=1, k \neq j}^{N} K_{cosine}(f^x_k, f^x_j)} \in [0, 1], \tag{3}$$

and

$$g_{i|j} = \frac{K_{cosine}(f^y_i, f^y_j)}{\sum_{k=1, k \neq j}^{N} K_{cosine}(f^y_k, f^y_j)} \in [0, 1], \tag{4}$$

where the cosine kernel function K_{cosine} is employed for estimating the probabilty distribution, formulated with two vectors a and b as:

$$K_{cosine}(a, b) = \frac{1}{2}(\frac{a^\top b}{||a||_2 ||b||_2} + 1) \in [0, 1]. \tag{5}$$

As Turlach et al. [56] suggested, this kernel function avoids the bandwidth choosing in Gaussian kernel, and it boosts performance compared with the Euclidean measures as Wang et al. [57] suggested.

To minimize the difference of two estimated probability distributions, we avail the Kullback-Leibler (KL) divergence as the similarity metric, formulated as:

$$D_{KL}(\mathcal{G}'||\mathcal{G}) = \int_{\mathbf{t}} \mathcal{G}'(\mathbf{t}_x) \log \frac{\mathcal{G}'(\mathbf{t}_x)}{\mathcal{G}(\mathbf{t}_y)} d\mathbf{t}. \tag{6}$$

where $\mathbf{t}_x \in \mathcal{X}$ and $\mathbf{t}_y \in \mathcal{Y}$. In practical implementation, we avail the mini-batch that consists of N samples for approximation, aiming for acceleration in a parallel fashion.
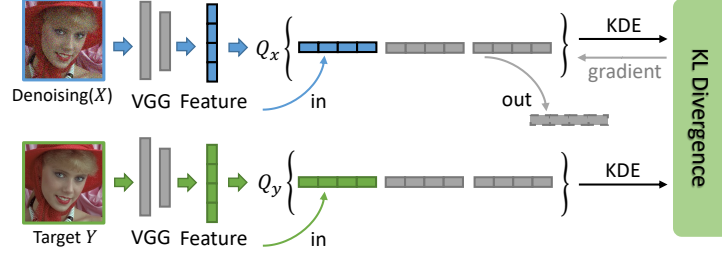
Fig. 3: **Sampling with Historic Gradients.** We approximate the divergence with historic sampling by using two queues to bypass the GPU memory limits.

3.2   Memorized Historic Sampling

Intuitively, the number of selected samples in a mini-batch should be as large as possible during training. However, in practical implementation, increasing the number of samples is greatly limited by the GPU memory. Such a limitation is more serious in the extracted semantic feature space, and hence greatly limits the effectiveness of our method.

To bypass the limitation, we introduce a Memorized Historic Sampling strategy, visualized in Figure 3. It maintains two *queues* of feature samples, *i.e.*, $Q^{\mathcal{X}}$ and $Q^{\mathcal{Y}}$ that can store historic features from previous mini-batches with limited GPU memory cost. In doing so, we can estimate the probability distributions among queues instead of mini-batches. Therefore, it allows a larger number of samples and a relatively smaller mini-batch used at runtime. The queue is updated according to the First-In-First-Out rule, which enforces the historical samples in the queue are always newest, and hence it allows the probability distribution to be more consistent with the immediate state. Based on such a strategy, we can formulate Equation 3 as below:

$$g'_{i|j} = \frac{K_{cosine}(Q_i^{\mathcal{X}}, Q_j^{\mathcal{X}})}{\sum_{k=1, k\neq j}^{q} K_{cosine}(Q_k^{\mathcal{X}}, Q_j^{\mathcal{X}})} \in [0,1], \tag{7}$$

and

$$Q_{1...N}^{\mathcal{X}}, Q_{N...q}^{\mathcal{X}} \leftarrow f_{\{1...N\}}^{x}, Q_{\{1...q-N\}}^{\mathcal{X}}. \tag{8}$$

where q is the queue size of the appied queue for extracted features f_x from a sigle mini-batch with the number of N, and $q \gg N$.

For example, the maximum size of a mini-batch can only be 32 in a single GPU card with a memory of 12GB, but the number of samples is usually set as 128 to ensure the accuracy of the estimation, which is not practical in a single GPU. By using the queue in the size of 128, we can directly use the current mini-batch with features from 3 historical memorized mini-batch to perform an estimation with 128 samples, while without using additional 12×3 GB memory at running time. It is because the queue that saves historical features costs less
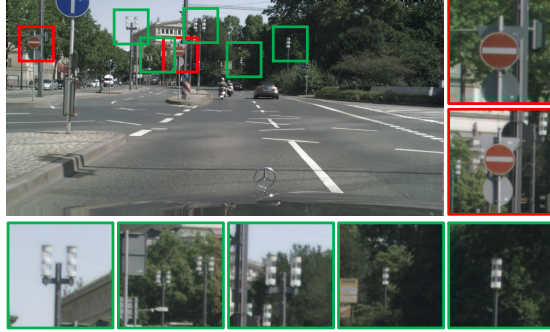
Fig. 4: **Sampling with Internal Patches.** Patches cropped from a single image may consist of different semantic objects that showed in different appearances.

GPU memory compared with the procedure of feature extraction. Similar strategies for enlarging the number of samples also exist, *e.g*, memory bank [60] and momentum encoder [14]. Compared with them, our memorized historic queue is simpler but also enlarges the maximum number of samples to be used without additional GPU memory. In the supplement we provide discussion about the effects of different queue sizes.

3.3   Patch-Wise Internal Probabilities

The most straightforward way to construct samples for the mini-batch is to randomly choose multiple images from the domains $\mathcal{X}$ and $\mathcal{Y}$, respectively. Even though it is elegant, it fails to exploit another crucial probability in the single image, *i.e.*, internal probability of patches from a single image, or internal statistics, which has been widely employed and empirically evaluated in many image restoration tasks [39, 48, 49, 66]. As illustrated in Figure 4, we can notice that the cropped patches from the single image, specifically the complex-semantic image, can be seen as multiple single-semantic images. In addition, the probability estimated on multiple complex-semantic images is not always accurate in some cases, *e.g.*, the denoising learning [62] in mixed noise levels. In such a case, conventional sampling results in an inaccurate estimation, because the restored images in the mini-batch come from different noisy levels. These restored images distribute in different manifolds in the semantic feature space. In contrast, by extracting patches from a single image resulted from extreme similar degradation, their similarity allows the probabilities to be accurately estimated.

In our method, a sliding window in a spatial size of $K \times K$ is availed to extract patches from the restored image $\tilde{Y}$ and clear image Y. These patches are then inputted into the high-level vision network $\Phi(\cdot)$ as a single mini-batch and transformed into feature sets, which can be formulated as:

$$f^x = \Phi(\text{sliding\_window}(\tilde{Y})).\tag{9}$$

Then we exploit Equation 7 to estimate the probability distribution in the feature sets respectively for divergence approximation. In the empirical evaluation, shown in Table 1, though the vanilla version achieves gains on the restoration performance, the generated images contain less semantic details (lower MIoU performance). This decreasing may boil down to incorrect probability estimation. In contrast, with the help of internal probability, our restoration network achieves a performance leap in both the restoration and semantic evaluation.

4   Experiments

Different from conventional denoising works, we focus on not only the restoration performance, but also the semantic accuracy, *i.e.*, how the denoised images can be understood by semantic segmentation networks, as well as whether the method can be extended into the other restoration tasks. Such a similar evaluation protocol is also employed in recent restoration works [21, 23, 52]. Therefore, our experiments are divided into three parts, including *Cityscape Denoising and Segmentation* [7], *Face Super-resolution and Alignment* [32], and *Natural Image Restoration* [2]. More details please refer to the supplemental.

4.1   Cityscape Denoising and Segmentation

To demonstrate the superiority of our method, we conduct complementary denoising and segmentation experiments on the Cityscapes dataset. The most representative denoising network, *i.e.*, FFDNet [62], as well as the state-of-the-art denoising networks, *i.e.*, CBDNet [13] and SADNet [4] are availed as the generation network $G(\cdot)$. Various objectives are applied, *i.e.*, $\mathcal{L}_1$, $\mathcal{L}_{SSIM}$ [59], $\mathcal{L}_{Perceptual}$ [17], $\mathcal{L}_{LPIPS}$ [63], $\mathcal{L}_{Contextual}$ [36], $\mathcal{L}_{CrossEntropy}$ [28], and ours. Notably, $\mathcal{L}_{CrossEntropy}$ is conducted with the HRNet48 that pre-trained on the Cityscapes dataset, which requires semantic labels during the denoising learning. In contrast, ours does not need any additional data. We then modify the original loss function of CBDNet and SADNet, *i.e.*, $\mathcal{L}_2 + \mathcal{L}_{Asymmetric} + \mathcal{L}_{TV}$ and $\mathcal{L}_2$, by attaching our proposed objective. For convenience, we set the size of the sliding window K as 224 and its stride as 56.

For denoising training, we construct noisy images by adding additive color Gaussian noise of noise level $\sigma \in [0, 75]$ to the clean images from the Cityscapes training set. The images are randomly cropped into 512×512 patches in a mini-batch size of 64. Other settings are kept the same as the settings in FFDNet. For evaluation, we first measure appearance similarities between restored images and corresponding clear images in the Cityscapes validation set, in noisy levels $\{25, 35, 50\}$, which is commonly selected by the denoising community. We then measure the semantic segmentation accuracy on restored images in the term of Mean Intersection-over-Union (MIoU) in 19 pre-defined semantic classes, i.e., *road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle*, and *bicycle*.

Table 1: **Quantitative performance comparison on the cityscape denoising and segmentation.** The comprasion is conducted with various state-of-the-art denoising objectives and ours on the representative denoising networks.

Method (Backbone)	Objective	Noise-Level $\sigma=25$			Noise-Level $\sigma=35$			Noise-Level $\sigma=50$		
		PSNR ↑	SSIM ↑	MIoU (%) ↑	PSNR ↑	SSIM ↑	MIoU (%) ↑	PSNR ↑	SSIM ↑	MIoU (%) ↑
FFDNet [62]	$\mathcal{L}_1$	$35.033^{(6)}$	$0.925^{(6)}$	$0.605^{(8)}$	$34.074^{(6)}$	$0.912^{(6)}$	$0.537^{(8)}$	$32.845^{(6)}$	$0.895^{(6)}$	$0.451^{(7)}$
	$+ \mathcal{L}_{SSIM}$ [59]	$35.567^{(3)}$	$0.935^{(2)}$	$0.642^{(2)}$	$34.469^{(4)}$	$0.922^{(2)}$	$0.584^{(2)}$	$33.180^{(3)}$	$0.906^{(2)}$	$0.450^{(8)}$
	$+ \mathcal{L}_{Perceptual}$ [17]	$34.319^{(7)}$	$0.912^{(7)}$	$0.629^{(4)}$	$33.486^{(7)}$	$0.899^{(7)}$	$0.582^{(4)}$	$32.383^{(7)}$	$0.881^{(7)}$	$0.509^{(2)}$
	$+ \mathcal{L}_{LPIPS}$ [63]	$35.551^{(4)}$	$0.929^{(4)}$	$0.613^{(6)}$	$34.463^{(5)}$	$0.916^{(4)}$	$0.541^{(7)}$	$33.138^{(5)}$	$0.899^{(4)}$	$0.452^{(6)}$
	$+ \mathcal{L}_{Contextual}$ [36]	$25.115^{(8)}$	$0.762^{(8)}$	$0.628^{(5)}$	$24.938^{(8)}$	$0.758^{(8)}$	$0.583^{(3)}$	$24.775^{(8)}$	$0.753^{(8)}$	$0.509^{(2)}$
	$+ \mathcal{L}_{CrossEntropy}$ [28]	$35.913^{(2)}$	$0.932^{(3)}$	$0.630^{(3)}$	$34.800^{(2)}$	$0.919^{(3)}$	$0.565^{(5)}$	$33.477^{(2)}$	$0.903^{(3)}$	$0.491^{(4)}$
D2SM (Ours)	w/o. Internal	$35.543^{(5)}$	$0.929^{(4)}$	$0.612^{(7)}$	$34.475^{(3)}$	$0.916^{(4)}$	$0.546^{(6)}$	$33.167^{(4)}$	$0.899^{(4)}$	$0.463^{(5)}$
	w/. Internal	$\mathbf{36.454}^{(1)}$	$\mathbf{0.936}^{(1)}$	$\mathbf{0.644}^{(1)}$	$\mathbf{35.206}^{(1)}$	$\mathbf{0.923}^{(1)}$	$\mathbf{0.587}^{(1)}$	$\mathbf{33.807}^{(1)}$	$\mathbf{0.907}^{(1)}$	$\mathbf{0.520}^{(1)}$
CBDNet [13]	-	$36.152^{(3)}$	$0.936^{(2)}$	$0.655^{(3)}$	$34.964^{(3)}$	$0.923^{(3)}$	$0.599^{(3)}$	$33.613^{(3)}$	$0.907^{(3)}$	$0.539^{(3)}$
	w/o. Internal	$36.254^{(2)}$	$0.935^{(3)}$	$0.679^{(2)}$	$35.158^{(2)}$	$0.925^{(2)}$	$0.631^{(2)}$	$33.904^{(2)}$	$0.911^{(2)}$	$0.550^{(2)}$
	w/. Internal	$\mathbf{36.899}^{(1)}$	$\mathbf{0.941}^{(1)}$	$\mathbf{0.691}^{(1)}$	$\mathbf{35.596}^{(1)}$	$\mathbf{0.929}^{(1)}$	$\mathbf{0.652}^{(1)}$	$\mathbf{34.172}^{(1)}$	$\mathbf{0.914}^{(1)}$	$\mathbf{0.600}^{(1)}$
SADNet [4]	-	$36.310^{(3)}$	$0.936^{(3)}$	$0.674^{(3)}$	$35.081^{(3)}$	$0.924^{(3)}$	$0.637^{(3)}$	$33.730^{(3)}$	$0.908^{(3)}$	$0.581^{(3)}$
	w/o. Internal	$36.822^{(2)}$	$0.940^{(2)}$	$0.691^{(2)}$	$35.247^{(2)}$	$0.924^{(2)}$	$0.655^{(2)}$	$34.133^{(2)}$	$0.912^{(2)}$	$0.600^{(2)}$
	w/. Internal	$\mathbf{37.130}^{(1)}$	$\mathbf{0.943}^{(1)}$	$\mathbf{0.701}^{(1)}$	$\mathbf{35.839}^{(1)}$	$\mathbf{0.931}^{(1)}$	$\mathbf{0.670}^{(1)}$	$\mathbf{34.440}^{(1)}$	$\mathbf{0.916}^{(1)}$	$\mathbf{0.634}^{(1)}$

Quantitative Comparison. In Table 1, it is easy to see that ours outperforms all compared objectives largely in PSNR, SSIM, and MIoU metrics on all noisy levels, when applied in the same backbone. Compared with the state-of-the-art objectives that combine high-level vision tasks, *i.e.*, $\mathcal{L}_{CrossEntropy}$ [28], which requires the semantic label of images during training, ours still performs it by 0.542dB in PSNR, 1.4% in MIoU, without using any additional data. Besides, ours shows strong robustness when adopted with different network architectures and objectives, *e.g.*, it helps the original CBDNet improves 0.747dB in PSNR and 0.5% in MIoU. Also, the comparison between using and without using internal probability further demonstrates its superiority for complex-semantic images.

Qualitative Comparison. Though $\mathcal{L}_{Perceptual}$ applied in restoration methods has been proven to lead to better perceptual quality in restored images, we find that ours significantly outperforms it with more visually pleasant and exact details as shwon in Figure 5. As shown in Figure 6, our restored results best preserve the edge of the character "**S**" in the red rectangle area. Besides, the blue rectangle area shows the best sharp details in our restored results compared with others. With regard to the segmentation evaluation, restored results from restoration networks trained with ours can best be segmented accurately. For instance, as two green rectangle areas are shown in the Figure 6, our result is the only one that is successfully recognized into *traffic light*. This indicates that ours can best preserve semantic details during restoration in the way of divergence minimization.

Distribution Visualization. In order to get insights into the probability distribution, here we visualize the semantic feature space estimated by restored images and clear images. To elaborate, we randomly select 500 animal images that belong to 10 categories, *i.e.*, *cat, dog, chicken, cow, horse, sheep, squirrel, elephant, butterfly*, and *spider*. We then process their noisy version (*i.e.* adding additive color Gaussian noise of noise level $\sigma=25$) with the FFDNet pre-trained on the noisy Cityscapes dataset. After that, the denoised images, as well as the clear images, are inputted into the pre-trained ResNet101 [16] to extract semantic feature maps. As such, we can visualize the distribution of semantic features

Noisy Image $\mathcal{L}_1$ $\mathcal{L}_{SSIM}$ [59] $\mathcal{L}_{Perceptual}$ [17] $\mathcal{L}_{LPIPS}$ [63]

$\mathcal{L}_{Contextual}$ [36] $\mathcal{L}_{CrossEntropy}$ [28] Ours w/o. Internal   Ours   Clear Image
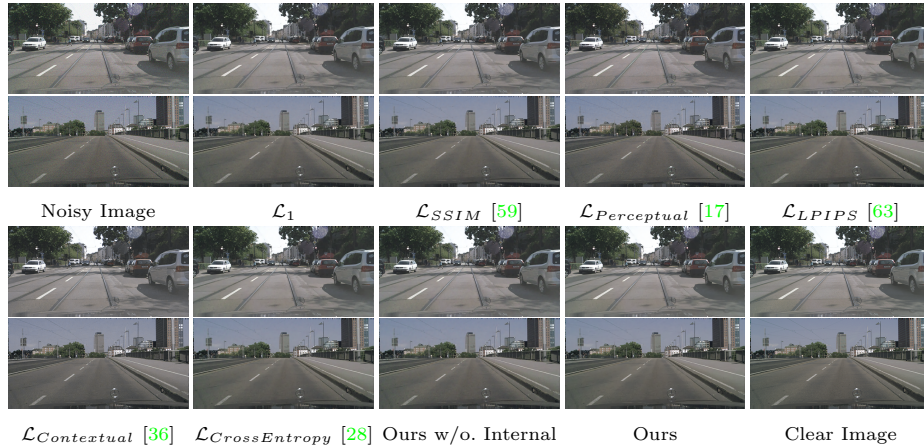
Fig. 5: Qualitative comparison on the denoising results. Ours results contain the most fine-grained high-frequency information and more visual pleasant details. (400% Zoom is recommended to see their difference in details and color bias.)

with the t-SNE [34] in 2D coordinates. Compared with others, the visualized distribution from our restored images best preserves the distribution of clear images in the semantic feature space. This indicates that our proposed method indeed implicitly minimizes the probability distribution divergence between restored images and clear images in the semantic feature space.

4.2   Face Super-resolution and Alignment

Here we extend D2SM with historic sampling as the objective and conduct face super-resolution learning under the settings of DICNet [33]. For the evaluation of high-level vision applications, we exploit the face alignment as a measurement, and its accuracy is denoted in the term of NRMSE. The queue size is set as 256 and the mini-batch size is 32, which means the probability is estimated among 32 face images instead interal patches.

In table 2, we show the quantitative performance comparison between DIC-Net and ours, as well as the other state-of-the-art methods. Notably, the face alignment network is pre-trained on the CelebA dataset, and hence its evaluation on the Helen dataset is generally not good enough, which can only be used for reference. By combining our objective with existing $\mathcal{L}_1$ and $\mathcal{L}_{Alignment}$ proposed by DICNet, our modified version successfully outperforms the original DICNet and DICNet in the GAN manner (DICGAN), in both the distortion measurement and alignment measurement. In contrast, the original DICNet can only achieve leading performance in the distortion measurement but is poor in alignment measurement, while the DICGAN can only achieve leading performance in the alignment measurement but is bad at distortion measurement.
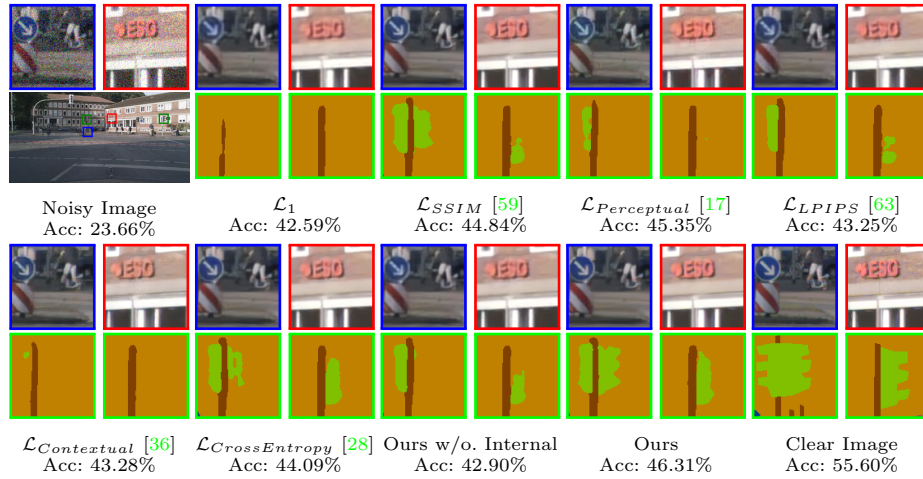
Fig. 6: Qualitative comparison on the denoising and segmentation results. Ours preserves most of the semantic details, including the human shape and font edge in the highlighted area. Additionally, in the shown segmentation results, our result is the only one that can be successfully recognized into *traffic light*.

4.3   Natural Image Restoration

Different from cityscape images collected from limited scenes, natural images contain more diverse and complex semantics. Therefore, the intrinsic semantics of natural images are more complex and diverse, which indicate a more challenging probability distribution estimation for our method. To validate our effectiveness in such cases, here we follow the settings of the state-of-the-art dehazing method, *i.e.*, MSBDN-DFF [10] on the end-to-end dehazing tasks [37], and we extend the dehazing network with our proposed objective. As the quantitative performance comparison shown in Table 3, though the density estimation is challenging, our method can successfully outperform the compared method in both the indoor scenes and outdoor scenes without additional cost. In Figure 7, we show some randomly highlighted visual results for comparison, and all of our results contain the most clear appearance with less haze remained. Specifically, we can notice that objects, *e.g.*, *floor*, *chair*, *roof*, *toy* that contain certain semantics, are better restored with accurate color than the method trained with pixel-wise loss functions only. This phenomenon further demonstrates the semantics transferring ability of our method, which regularizes restored objects to be semantic consistent and avoids incorrect color that against its semantics.

5   Discussion

Designing Choice of KL Divergence. The way of the divergence estimation between $\mathcal{G}'$ and $\mathcal{G}$ accounts a lot in our method. We empirically employed the Kullback-Leibler (KL) divergence weighted by $\mathcal{G}'$ for D2SM. Indeed, there are

Table 2: Quantitative performance comparison on the face super-resolution and alignment. By simply attaching our objective into the DICNet, our method can outperform the state-of-the-art DICNet and DICGAN in both the distortion measurement and high-level vision application measurement.

Method (Backbone)	Objective	CelebA Dataset			Helen Dataset		
		PSNR ↑	SSIM ↑	NRMSE ↓	PSNR ↑	SSIM ↑	NRMSE ↓
Bicubic	-	$23.58^{(10)}$	$0.6285^{(10)}$	$0.3385^{(8)}$	$23.89^{(10)}$	$0.6751^{(10)}$	$0.4577^{(8)}*$
SRResNet [20] [CVPR-17]	$\mathcal{L}_2$	$25.82^{(6)}$	$0.7369^{(6)}$	-	$25.30^{(6)}$	$0.7297^{(7)}$	-
URDGN [61] [ECCV-16]	$\mathcal{L}_2 + \mathcal{L}_{GAN}$	$24.63^{(8)}$	$0.6851^{(9)}$	-	$24.22^{(9)}$	$0.6909^{(9)}$	-
RDN [65] [ECCV-18]	$\mathcal{L}_1$	$26.13^{(5)}$	$0.7412^{(5)}$	$0.1415^{(4)}$	$25.34^{(5)}$	$0.7249^{(8)}$	$0.4437^{(7)}*$
PFSR [18] [BMVC-18]	$\mathcal{L}_2 + \mathcal{L}_{Perceptual} + \mathcal{L}_{GAN} + \mathcal{L}_{Heatmap} + \mathcal{L}_{Attention}$	$24.43^{(9)}$	$0.6991^{(8)}$	$0.1917^{(7)}$	$24.73^{(8)}$	$0.7323^{(6)}$	$0.3498^{(4)}*$
FSRNet [6] [CVPR-18]	$\mathcal{L}_2 + \mathcal{L}_{Perceptual}$	$26.48^{(3)}$	$0.7718^{(3)}$	$0.1430^{(5)}$	$25.90^{(4)}$	$0.7759^{(3)}$	$0.3723^{(6)}*$
	$\mathcal{L}_2 + \mathcal{L}_{Perceptual} + \mathcal{L}_{GAN}$	$25.06^{(7)}$	$0.7311^{(7)}$	$0.1463^{(6)}$	$24.99^{(7)}$	$0.7424^{(5)}$	$0.3408^{(3)}*$
DICNet [33] [CVPR-20]	$\mathcal{L}_1 + \mathcal{L}_{Alignment}$	$27.28^{(2)}$	$0.7929^{(2)}$	$0.1345^{(3)}$	$26.69^{(2)}$	$0.7933^{(2)}$	$0.3674^{(5)}*$
	$\mathcal{L}_1 + \mathcal{L}_{Alignment} + \mathcal{L}_{Perceptual} + \mathcal{L}_{GAN}$	$26.34^{(4)}$	$0.7562^{(4)}$	$0.1319^{(2)}$	$25.96^{(3)}$	$0.7624^{(4)}$	$\mathbf{0.3336}^{(1)}*$
Ours	w/o. Internal	$\mathbf{27.39}^{(1)}$	$\mathbf{0.7973}^{(1)}$	$\mathbf{0.1292}^{(1)}$	$\mathbf{26.94}^{(1)}$	$\mathbf{0.8005}^{(1)}$	$0.3366^{(2)}*$

Table 3: Quantitative comparison on the natural image dehazing. Our proposed objective is capable of being extended to the dehazing task based on MSBDN-DFF, which shows superiority in both the indoor and outdoor datasets.

Method	Metric	DCP [15]	MSCNN [44]	DcGAN [25]	GFN [45]	PFFNet [37]	GDN [29]	DuRN [30]	MSBDN-DFF [10]	Ours
I-HAZE	PSNR↑	$14.43^{(9)}$	$15.22^{(8)}$	$16.06^{(5)}$	$15.84^{(7)}$	$16.01^{(6)}$	$16.62^{(4)}$	$21.23^{(3)}$	$23.93^{(2)}$	$\mathbf{24.31}^{(1)}$
	SSIM↑	$0.752^{(6)}$	$0.755^{(5)}$	$0.733^{(9)}$	$0.751^{(7)}$	$0.740^{(8)}$	$0.787^{(4)}$	$0.891^{(2)}$	$\mathbf{0.902}^{(1)}$	
O-HAZE	PSNR↑	$16.78^{(9)}$	$17.56^{(8)}$	$19.34^{(4)}$	$18.16^{(7)}$	$18.76^{(6)}$	$18.92^{(5)}$	$20.45^{(3)}$	$24.36^{(2)}$	$\mathbf{24.79}^{(1)}$
	SSIM↑	$0.653^{(8)}$	$0.650^{(9)}$	$0.681^{(4)}$	$0.671^{(6)}$	$0.669^{(7)}$	$0.672^{(5)}$	$0.688^{(3)}$	$0.749^{(2)}$	$\mathbf{0.787}^{(1)}$

some other ways to estimate the divergence. The most similar one, *i.e.*, inverse KL divergence weighted by $\mathcal{G}$, which is also asymmetrical. Based on KL divergence, another way to estimate the divergence is Jensen–Shannon divergence, which is symmetric and can be seen as the smoothed version of KL divergence, formulated as:

$$D_{JS}(\mathcal{G}'||\mathcal{G}) = \frac{1}{2}D_{KL}(\mathcal{G}'||\mathcal{G}) + \frac{1}{2}D_{KL}(\mathcal{G}||\mathcal{G}'). \tag{10}$$

According to [3], the optimization procedure of the optimal discriminator $D*$ in GAN yields minimizing the JS divergence, formulated as:

$$\mathcal{L}(D*, G) = 2D_{JS}(\mathcal{G}'||\mathcal{G}) - 2\log 2. \tag{11}$$

Method	σ	PSNR ↑	SSIM ↑	MIoU (%) ↑
FFDNet	25	35.03	0.925	0.605
$+ \mathcal{L}_{iKLD}$	25	35.97	0.931	0.638
$+ \mathcal{L}_{JSD}$	25	36.31	0.935	0.640
$+ \mathcal{L}_{GAN}$	25	35.55	0.931	0.621
Ours	25	**36.45**	**0.936**	**0.644**

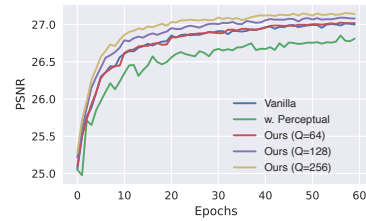Table 4: Performance comparison with different distribution divergence.

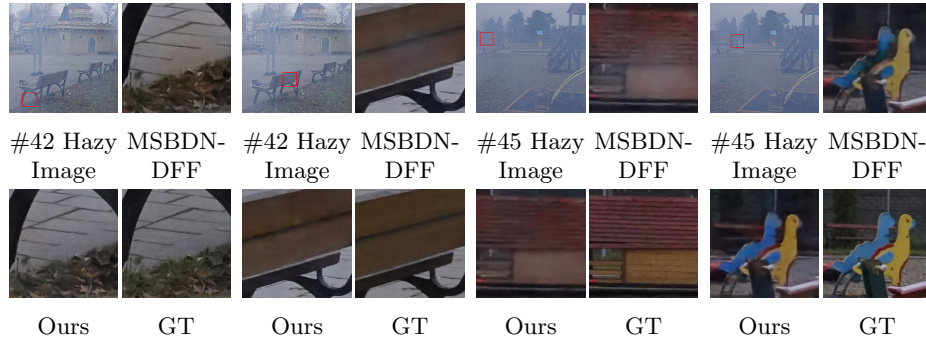Fig. 8: Convergence visualization between different queue size.

Fig. 7: Qualitative comparison on the real-world dehazing. Compared with the SOTA method that employs pixel-wise loss functions, our extended version better recover the scenes under severe ill-posed distortion.

Here we present the quantitative comparison with the three additional divergence estimation or objectives in Cityscapes. The comprasion empirically proves our superiority of using KL-divergence compared with the others in the denoising tasks.

Designing Choice of Queue Size. The convergence curve visualized in the Figure 8 further demonstrates that our proposed method significantly accelerates convergence with the proposed memorized historic sampling. From the figure we can notice that the applied historic sampling with large queue size $(Q > 64)$ can greately accelerate the learning, while the vanilla version can only achieve poor performance $(Q \leq 64)$. Thus, in our practical implementation, we chose the queue size with the possible max value under the computational limitation.

6   Conclusion

We propose a simple but practical method for facilitating the restoration learning and preserving the semantic attribute. It does not rely on any external information nor introduce additional parameters. By implicitly approximating the divergence on the semantic feature space, we can force existed generation networks to learn to preserve semantic attributes during restoration learning. We further transfer the method from the single-semantic image to the complex-semantic image *i.e.* natural image by using internal statistics. Empirically evaluation validates that the proposed method can be adapted to various restoration tasks and network architectures with general performance improvement.

Acknowledgments

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPR Workshops (2017) 4
2. Ancuti, C., Ancuti, C.O., Timofte, R., De Vleeschouwer, C.: I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images. In: ICACIVS (2018) 9
3. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862 (2017) 13
4. Chang, M., Li, Q., Feng, H., Xu, Z.: Spatial-Adaptive Network for Single Image Denoising. In: ECCV (2020) 9, 10
5. Chen, H., Wang, Y., Xu, C., Xu, C., Tao, D.: Learning student networks via feature embedding. TNNLS (2020) 5, 6
6. Chen, Y., Tai, Y., Liu, X., Shen, C., Yang, J.: Fsrnet: End-to-end learning face super-resolution with facial priors. In: CVPR (2018) 13
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016) 2, 4, 9
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 1, 2, 4
9. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. TPAMI (2015) 1
10. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.H.: Multi-Scale Boosted Dehazing Network With Dense Feature Fusion. In: CVPR (2020) 12, 13
11. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: NeurIPS (2016) 4, 5
12. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks. In: NeurIPS (2014) 1
13. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward Convolutional Blind Denoising of Real Photographs. In: CVPR (2019) 9, 10
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum Contrast for Unsupervised Visual Representation Learning. In: CVPR (2020) 3, 8
15. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. TPAMI (2010) 13
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: CVPR (2016) 10
17. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016) 1, 4, 5, 9, 10, 11, 12
18. Kim, D., Kim, M., Kwon, G., Kim, D.S.: Progressive face super-resolution via attention to facial landmark. In: BMVC (2019) 13
19. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: ECCV (2012) 4
20. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017) 4, 5, 13
21. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: Aod-net: All-in-one dehazing network. In: ICCV (2017) 9
22. Li, C.L., Chang, W.C., Cheng, Y., Yang, Y., Póczos, B.: Mmd gan: Towards deeper understanding of moment matching network. arXiv preprint arXiv:1705.08584 (2017) 2

23. Li, J., Fang, F., Li, J., Mei, K., Zhang, G.: MDCN: Multi-scale Dense Cross Network for Image Super-Resolution. TCSVT (2020) 9
24. Li, J., Fang, F., Mei, K., Zhang, G.: Multi-scale residual network for image super-resolution. In: ECCV (2018) 1
25. Li, R., Pan, J., Li, Z., Tang, J.: Single image dehazing via conditional generative adversarial network. In: CVPR (2018) 13
26. Li, Y., Swersky, K., Zemel, R.: Generative moment matching networks. In: ICML (2015) 4
27. Liu, D.: Connecting Low-Level Image Processing and High-Level Vision via Deep Learning. In: IJCAI (2018) 4
28. Liu, D., Wen, B., Jiao, J., Liu, X., Wang, Z., Huang, T.S.: Connecting image denoising and high-level vision tasks via deep learning. TIP (2020) 4, 9, 10, 11, 12
29. Liu, X., Ma, Y., Shi, Z., Chen, J.: Griddehazenet: Attention-based multi-scale network for image dehazing. In: ICCV (2019) 13
30. Liu, X., Suganuma, M., Sun, Z., Okatani, T.: Dual residual networks leveraging the potential of paired operations for image restoration. In: CVPR (2019) 13
31. Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., Duan, Y.: Knowledge distillation via instance relationship graph. In: CVPR (2019) 5, 6
32. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015) 4, 9
33. Ma, C., Jiang, Z., Rao, Y., Lu, J., Zhou, J.: Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In: CVPR (2020) 11, 13
34. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. JMLR (2008) 2, 6, 11
35. Mechrez, R., Talmi, I., Shama, F., Zelnik-Manor, L.: Maintaining natural image statistics with the contextual loss. In: ACCV (2018) 1, 4, 5
36. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: ECCV (2018) 4, 5, 9, 10, 11, 12
37. Mei, K., Jiang, A., Li, J., Wang, M.: Progressive feature fusion network for realistic image dehazing. In: ACCV (2018) 12, 13
38. Mei, K., Ye, S., Huang, R.: Sdan: Squared deformable alignment network for learning misaligned optical zoom. In: ICME (2021) 4
39. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV (2020) 8
40. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR (2019) 5, 6
41. Passalis, N., Tefas, A.: Learning deep representations with probabilistic knowledge transfer. In: ECCV (2018) 5, 6
42. Passalis, N., Tzelepi, M., Tefas, A.: Probabilistic Knowledge Transfer for Lightweight Deep Representation Learning. TNNLS (2020) 5, 6
43. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: ICCV (2019) 5, 6
44. Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., Yang, M.H.: Single image dehazing via multi-scale convolutional neural networks. In: ECCV (2016) 13
45. Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., Yang, M.H.: Gated fusion network for single image dehazing. In: CVPR (2018) 13
46. Santos, C.N.d., Mroueh, Y., Padhi, I., Dognin, P.: Learning implicit generative models by matching perceptual features. In: CVPR (2019) 4
47. Scott, D.W.: Multivariate density estimation: theory, practice, and visualization (2015) 3

48. Shaham, T.R., Dekel, T., Michaeli, T.: Singan: Learning a generative model from a single natural image. In: ICCV (2019) 8
49. Shocher, A., Cohen, N., Irani, M.: Zero-Shot Super-Resolution Using Deep Internal Learning. In: CVPR (2018) 8
50. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015) 1, 4
51. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: ECCV (2016) 4
52. Tian, Y., Krishnan, D., Isola, P.: Contrastive Representation Distillation. In: ICLR (2019) 9
53. Tolstikhin, I.O., Sriperumbudur, B.K., Schölkopf, B.: Minimax estimation of maximum mean discrepancy with radial kernels. In: NeurIPS (2016) 4
54. Torkkola, K.: Feature extraction by non-parametric mutual information maximization. JMLR (2003) 5
55. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: ICCV (2019) 5
56. Turlach, B.: Bandwidth Selection in Kernel Density Estimation: A Review. CORE and Institut de Statistique (1999) 6
57. Wang, D., Lu, H., Bo, C.: Visual tracking via weighted local cosine similarity. TCYB (2014) 6
58. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: ECCV Workshops (2018) 4
59. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004) 4, 9, 10, 11, 12
60. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised Feature Learning via Nonparametric Instance Discrimination. In: CVPR (2018) 3, 8
61. Yu, X., Porikli, F.: Ultra-resolving face images by discriminative generative networks. In: ECCV (2016) 13
62. Zhang, K., Zuo, W., Zhang, L.: FFDNet: Toward a fast and flexible solution for CNN-based image denoising. TIP (2018) 1, 8, 9, 10
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 1, 4, 9, 10, 11, 12
64. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to Learn, Learn to Zoom. In: CVPR (2019) 4
65. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018) 13
66. Zontak, M., Irani, M.: Internal statistics of a single natural image. In: CVPR (2011) 3, 4, 8