Deep Semantic Statistics Matching Denoising Network

Kangfu Mei Vishal M. Patel Rui Huang

Johns Hopkins University The Chinese University of Hong Kong, Shenzhen

ECCV TEL AVIV 2022

TL;DR

A new framework for generalizing image restoration networks on high-level vision tasks.

```
!pip install git+https://github.com/MKFMIKU/d2sm.git

from d2sm import DeST
dest = DeST("54", args.patch_size, args.patch_size, Q=args.dest_q, K=args.dest_k, mean_shift=True)

dest_loss = dest(self.hr, self.denoised_hr)
```

Problem Settings

By maximizing the correspondence between each pair of the CNN-restored results and the clear image, a CNN is trained to restore images from the degraded domain into the clear domain.

Recent work called perceptual loss is shown to be able to lead to better visual quality by maximizing the correspondence in the semantic feature space of pre-trained classification network.

GANs, which employ a discriminator to implicitly enforce the distribution of restored images to be consistent with the distribution of clear images in terms of KL- and JS- divergence.

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?

In this work, we exploit semantic features of pretrained classification networks and implicitly matches the probabilistic distribution of clear images at the semantic feature space.

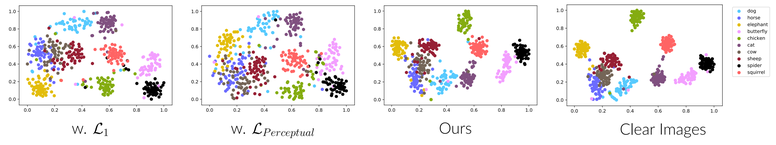

Figure 1. **t-SNE of denoised Images in the Semantic Feature Space.** Ours preserves most semantics like GT.

Contributions

- A new image restoration framework that minimizes the distribution divergence instead of the sample-to-sample distance in the semantic feature space.
- A new patch-wise fashion that can decompose complex semantics of images for efficient distribution approximation.
- The method substantially outperforms the original perceptual loss and other SOTA losses, especially in high-level vision tasks that validates D2SM indeed preserve semantics.

References

[1] Meng Chang, Qi Li, Huajun Feng, and Zhihai Xu. Spatial-Adaptive Network for Single Image Denoising. In *ECCV*, 2020.

[2] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward Convolutional Blind Denoising of Real Photographs. In *CVPR*, 2019.

[3] Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *TIP*, 2018.

Method

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 of them in the feature space of $\Phi(G(\cdot))$ (e.g. pretrained VGG-19 network) into the restoration learning.

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 $\mathcal{Y}$ as

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

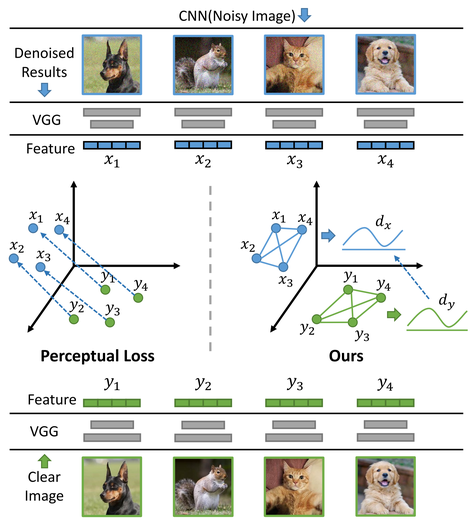

Figure 2. **Perceptual loss vs. Ours.**

To bypass the memory limitation, we introduce a Memorized Historic Sampling strategy by maintaining 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 as

$$g'_{i|j} = \frac{K_{cosine}(Q_i^{\mathcal{X}}, Q_j^{\mathcal{X}})}{\sum_{k=1,k\neq j}^{q} K_{cosine}(Q_k^{\mathcal{X}}, Q_j^{\mathcal{X}})} \in [0, 1], \tag{2}$$

and

$$Q^{\mathcal{X}}_{1\ldots N}, Q^{\mathcal{X}}_{N\ldots q} \leftarrow f^{x}_{\{1\ldots N\}}, Q^{\mathcal{X}}_{\{1\ldots q-N\}}. \tag{3}$$

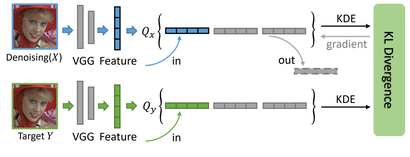

Figure 3. **Sampling with Historic Gradients.**

Results

Cityscape Denoising and Segmentation

We conduct complementary denoising and segmentation experiments on the Cityscapes dataset. 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.

Table 1. **Quantitative performance comparison on the cityscape denoising and segmentation.**

Method (Backbone)	Objective	Noise-Level σ=25 PSNR ↑	SSIM ↑	MIoU (%) ↑	Noise-Level σ=35 PSNR ↑	SSIM ↑	MIoU (%) ↑	Noise-Level σ=50 PSNR ↑	SSIM ↑	MIoU (%) ↑
FFDNet [3]	$\mathcal{L}_1$	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$^{(7)}$
	+ $\mathcal{L}_{SSIM}$	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$^{(8)}$
	+ $\mathcal{L}_{Perceptual}$	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$^{(2)}$
	+ $\mathcal{L}_{LPIPS}$	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$^{(6)}$
	+ $\mathcal{L}_{Contextual}$	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$^{(2)}$
	+ $\mathcal{L}_{CrossEntropy}$	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	**36.454**$^{(1)}$	**0.936**$^{(1)}$	**0.644**$^{(1)}$	**35.206**$^{(1)}$	**0.923**$^{(1)}$	**0.587**$^{(1)}$	**33.807**$^{(1)}$	**0.907**$^{(1)}$	**0.520**$^{(1)}$
CBDNet [2]	-	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	**36.899**$^{(1)}$	**0.941**$^{(1)}$	**0.691**$^{(1)}$	**35.596**$^{(1)}$	**0.929**$^{(1)}$	**0.652**$^{(1)}$	**34.172**$^{(1)}$	**0.914**$^{(1)}$	**0.600**$^{(1)}$
SADNet [1]	-	36.310$^{(3)}$	0.936$^{(3)}$	0.674$^{(3)}$	35.081$^{(3)}$	0.924$^{(2)}$	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	**37.130**$^{(1)}$	**0.943**$^{(1)}$	**0.701**$^{(1)}$	**35.839**$^{(1)}$	**0.931**$^{(1)}$	**0.670**$^{(1)}$	**34.440**$^{(1)}$	**0.916**$^{(1)}$	**0.634**$^{(1)}$

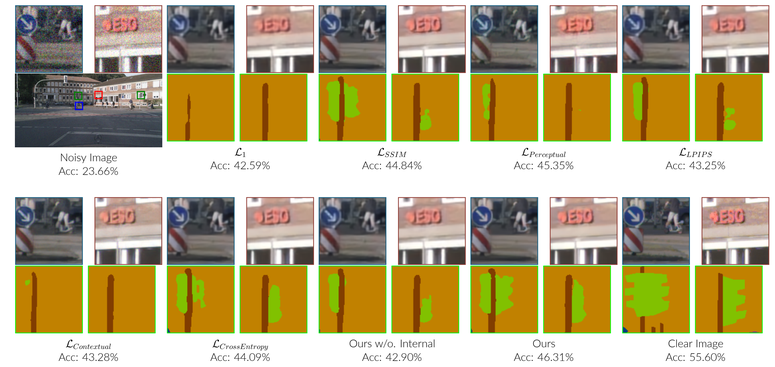

Figure 4. **Qualitative comparison on the denoising and segmentation results.** Ours preserves most of the semantic details, including the human shape and font edge in the highlighted area. Additionally, in the shown segmentation results, our result is the only one that can be successfully recognized into *traffic light*.

Results Analysis

Here we present the quantitative comparison with the three additional divergence estimation or objectives in Cityscapes. The convergence curve visualized below further demonstrates that our proposed method significantly accelerates convergence memorized historic sampling.

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 2. Performance comparison with different distribution divergence.

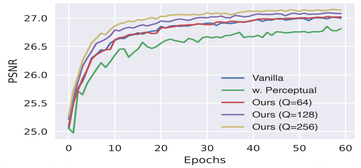

Figure 5. Convergence visualization between different queue size.