KERNEL ESTIMATION NETWORK FOR BLIND SUPER-RESOLUTION

Xiang Cao, Haibo Shen, Liangqi Zhang, Yihao Luo, Tianjiang Wang

School of Computer Science and Technology, Huazhong University of Science and Technology

ABSTRACT

Existing super-resolution (SR) methods commonly assume that the degradation kernels are fixed and known (e.g., bicubic downsampling or single Gaussian blurring kernel). However, these methods suffer a severe performance drop when the real degradations deviate from this assumption. To address this issue, this paper proposes a novel kernel estimation network (KENet) for kernel prediction. Specifically, KENet predicts the degradation kernels by optimizing the kernel space loss in a supervised way, without extra iterations at the inference time. Moreover, we introduce an adaptive attention loss to constrain the kernel optimization space, which can bias the allocation of trainable model parameters towards the most informative components of the estimation kernels. Extensive experiments on synthetic and real images show that the proposed KENet not only encourages a more accurate way to predict degradation kernels but also outperforms existing state-of-the-art blind SR methods when combined with non-blind SR methods.

Index Terms— Super-resolution, blind super-resolution, kernel estimation, degradation kernels

1. INTRODUCTION

Single image super-resolution (SISR) is an ill-posed problem for low-level vision tasks that aims to recover a high-resolution (HR) image with high-frequency details from its low-resolution (LR) counterpart. Since SRCNN [1] firstly attempted towards using a three-layer convolutional neural network for SISR, tremendous convolutional neural networks (CNNs) based methods have been proposed to improve super-resolution (SR) performance with the help of deeper networks [2, 3], residual blocks [4], and attention mechanisms [5]. The following degradation model obtains the training LR-HR image pairs of these methods:

$$I^{LR} = \left(I^{HR} \otimes \mathrm{k}\right) \downarrow_s \tag{1}$$

where I^{HR} and I^{LR} represent the HR and LR image, k is the degradation kernel. $\downarrow_s$ and $\otimes$ denote subsampling with a scale factor of s and linear convolution operation, respectively. However, most of these works commonly assume the degradation kernel k is fixed and known (e.g., bicubic downsampling or single Gaussian blurring kernel). Naturally, their performance may suffer from a significant drop when the real degradation differs from the assumption.

To address this problem, several non-blind SR methods [6, 7, 8, 9, 10] have been proposed, which take various degradation kernels as conditional inputs to establish the LR-HR mapping. For example, SRMD [7] concatenates the LR image and degradation maps as input to handle multiple degradations via training a single CNN model. USRNet [6] integrates the flexibility of model-based methods to handle multiple degradations by combining the LR image with scale factor, degradation kernels, and noise level as input. Although these non-blind models produce promising SR results when the ground truth degradation is known, the problem with unseen degradations limits their real-world applications.

The problem of SR with unknown degradation kernels is known as blind SR. In order to take advantage of the non-blind SR methods, the most common strategy [10, 11, 12, 13, 14] is to decompose blind SR into two sub-problems, i.e., kernel estimation and non-blind SR. Kernel estimation is a preliminary step of non-blind SR, which plays a crucial role in the final SR performance. Specifically, based on two "Deep-image-Prior" (DIP) networks [12], Double-DIP [14, 15] applies a joint optimization algorithm to capture the clean image and degradation kernel priors in the network parameter space. FKP [10] designs an invertible network to capture kernel distribution by learning a bijective mapping between kernel and latent space. In a different way, KernelGAN [11] uses an image-specific internal generative adversarial network (GAN) to estimate the degradation kernel. The generator of KernelGAN is a deep linear network, and the set of weights makes the estimated degradation kernel. However, these kernel estimation methods require thousands of self-training iterations at the inference time, which is time-consuming and unsuitable for real-world applications. Moreover, they predict the degradation kernel by minimizing the LR image reconstruction error is unsuitable, causing a suboptimal problem, where multiple degradation kernels could generate a similar LR image.

In this paper, we propose a novel kernel estimation network (KENet) that can predict a more accurate degradation

This work was supported in part by the National Natural Science Foundation of China under Grant 61572214 and Seed Foundation of Huazhong University of Science and Technology (2020kfyXGYJ114). (Corresponding author: Tianjiang Wang.)

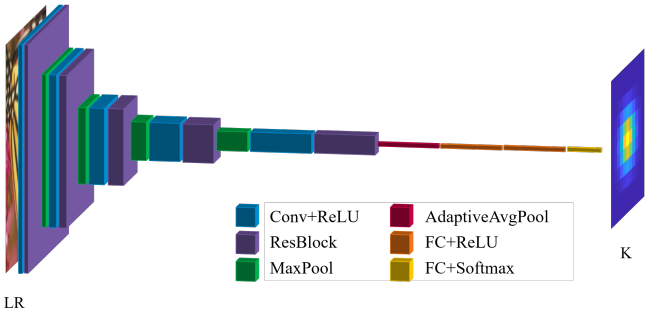

Fig. 1. Overview of the proposed kernel estimation network.

kernel without iterative optimization at inference time. Once trained, the proposed KENet can combine with existing state-of-the-art (SOTA) non-blind SR methods such as USRNet for blind SR, suitable for real-world applications that demand fast response. Instead of indirectly predicting the degradation kernel by minimizing the LR image reconstruction error, we find inherent degradation cues from the LR images by optimizing the kernel space loss in a supervised way. In addition, we propose a novel adaptive attention loss to constrain the kernel optimization space, which can be viewed as guidance to bias the allocation of trainable model parameters towards the most informative components of the estimation kernels. Extensive experiments show that the proposed KENet improves the accuracy of existing kernel estimation methods. Combined with the non-blind SR methods, it outperforms existing SOTA blind SR algorithms both on synthetic and real images.

2. METHOD

2.1. Motivation

As illustrated in Eq.1, the classical degradation model assumes that the LR image I^{LR} is obtained via a combination of blurring and downsampling from the HR image I^{HR}. Assume that the mapping $\mathcal{S}\left(I^{LR}, \mathrm{k}\right)$ is a well-trained non-blind SR model (e.g., USRNet [6]) with the degradation kernel k as input. As a preliminary step, the kernel estimation method is equivalent to finding the degradation kernel k that can combine with non-blind SR models to generate visually pleasant SR results for blind SR. Specifically, the most common strategy aims to estimate the degradation kernel by minimizing the LR image reconstruction error

$$\theta_{\mathcal{F}} = \arg\min_{\theta_{\mathcal{F}}} \left\| I^{LR} - \left(I^{HR} \otimes \mathcal{F}\left(I^{LR}; \theta_{\mathcal{F}}\right)\right) \downarrow_s \right\|_1 \quad (2)$$

where $\mathcal{F}$ is the kernel estimation model and $\theta_{\mathcal{F}}$ is the model parameter. However, there exist multiple degradation kernels that could generate similar LR images. In this case, minor reconstruction errors may lead to the wrong kernel prediction. Meanwhile, empirical and theoretical literature [16, 10] has demonstrated that the SR models would cause a severe performance drop with a slightly different downsampling kernel.

In view of this, we aim to find inherent degradation cues from the LR input directly in this paper. The operation can be express as

$$\theta_{\mathcal{F}} = \arg\min_{\theta_{\mathcal{F}}} \left\| \mathrm{k} - \mathcal{F}\left(I^{LR}; \theta_{\mathcal{F}}\right) \right\|_1 \quad (3)$$

2.2. Proposed Method

The overall framework of the proposed kernel estimation network is illustrated in Fig.1. Suppose the input LR image denotes as $I^{LR} \in \mathbb{R}^{H \times W \times C}$, where H and W represent the height and width of the input, C is the number of channels. The goal of the proposed kernel estimate network is to estimate the degradation kernel $\mathrm{k} \in \mathbb{R}^{l \times l}$ with the LR image as input, where $l \times l$ is the kernel size. In the beginning, a convolutional layer with ReLU activation and a residual block is applied to extract features and representations from the input LR image. Formally, the first feature maps f_0 are obtained as an operation

$$f_0 = (\mathrm{RB} \circ \mathrm{RL} \circ \mathrm{Conv})\left(I^{LR}\right) \quad (4)$$

where RL is ReLU activation, Conv and RB denote the convolutional and residual block operation, respectively. Then, a stack of a max-pooling operation, a convolution layer with ReLU activation, and a residual block are used to reduce the resolution of the features f_{i-1} and enlarge the receptive fields

$$f_i = (\mathrm{RB} \circ \mathrm{RL} \circ \mathrm{Conv} \circ \mathrm{MP})\left(f_{i-1}\right), i = 1, 2, 3 \quad (5)$$

where MP is the max pooling operation. The convolution layer in Eq.5 is used to expand feature maps and obtain more feature information. Finally, a global average pooling and three fully connected layers are applied to generate the degradation kernel:

$$\mathrm{k} = (\mathrm{SM} \circ \mathrm{FC} \circ \mathrm{RL} \circ \mathrm{FC} \circ \mathrm{RL} \circ \mathrm{FC} \circ \mathrm{AP})\left(f_3\right) \quad (6)$$

where AP and FC represent global average pooling and fully connected layer, respectively. SM denotes the Softmax activation, which ensures the non-negative and sum-to-one constraints on the degradation kernel.

All convolutional layers in KENet are of size 3×3 with $64\mathrm{s}$ output channels, where s represents the sth Conv operation. And each residual block includes two 3×3 convolutional layers with $64\mathrm{s}$ output channels and a ReLU activation in between. The output channels of each fully connected layer are 1024 except the final one, which is $l \times l$.

2.3. Adaptive Attention Loss

As depicted in the above section, estimating the degradation kernels by minimizing the LR reconstruction loss is not optimal, where multiple degradation kernels could generate similar LR images. Therefore we optimize kernel reconstruction

Table 1. Average PSNR[dB]/SSIM comparison of existing SR methods on the Set5 [17], Set14 [18], Bsd100 [19], and Urban100 [20] datasets.

Method	Scale	Set5	Set14	Bsd100	Urban100
Bicubic	X2	28.54/0.8571	26.41/0.7691	26.32/0.7156	23.74/0.7205
RCAN		28.71/0.8713	26.49/0.7835	26.58/0.7386	23.97/0.7462
KernelGAN+ZSSR		25.19/0.7693	24.62/0.7430	25.45/0.7411	22.98/0.7414
DAN		29.39/0.8880	27.18/0.8153	27.31/0.7883	24.84/0.8025
MZSR		27.96/0.8509	27.78/0.8131	27.49/0.8097	25.79/0.8210
KENet+USRNet(Ours)		**36.06/0.9503**	**32.11/0.8978**	**31.27/0.8789**	**29.15/0.8889**
GT+USRNet		36.37/0.9514	32.56/0.9021	31.47/0.8824	30.13/0.9023
Bicubic	X4	24.29/0.7261	23.15/0.6231	23.79/0.5979	21.03/0.5806
RCAN		23.70/0.7329	22.37/0.6185	22.71/0.5836	19.22/0.5333
KernelGAN+ZSSR		17.31/0.4388	17.19/0.4151	17.34/0.3879	15.96/0.4075
DAN		28.46/0.8475	26.10/0.7380	25.85/0.6975	23.22/0.764
MZSR		28.79/0.8543	26.59/0.7455	26.15/0.7086	23.98/0.7297
KENet+USRNet(Ours)		**31.23/0.8835**	**27.81/0.7705**	**27.04/0.7224**	**25.18/0.7650**
GT+USRNet		31.35/0.8851	27.90/0.7735	27.11/0.7260	25.32/0.7682

Table 2. Average PSNR[dB]/SSIM comparison of different loss functions on the Bsd100 [19] dataset.

Loss	Scale	Kernel PSNR	KENet+USRNet [6] PSNR/SSIM	Scale	Kernel PSNR	KENet+USRNet [6] PSNR/SSIM
$\mathcal{L}_1$	$\times 2$	59.52	31.20/0.8757	$\times 4$	46.11	26.95/0.7265
$\mathcal{L}_{att}$		61.51	31.27/0.8789		57.36	27.04/0.7224

loss to predict more accurate degradation kernels. Specifically, we introduce an adaptive attention loss to enhance the optimization to focus on the most informative components of the estimation kernels. Suppose p_{max} denotes the maximum pixel values of the ground truth kernel $\mathrm{k}_{gt} \in \mathbb{R}^{l \times l}$. For arbitrary pixel $\mathrm{k}^*(i,j)$ in estimation kernel $\mathrm{k}^* \in \mathbb{R}^{l \times l}$, we modify k^* based on p_{max}:

$$\mathrm{k}^*_{\lambda_t}(i,j) = \begin{cases} \mathrm{k}^*(i,j), & \text{if } \mathrm{k}_{gt}(i,j) > \lambda_t p_{max} \\ \mathrm{k}_{gt}(i,j), & \text{otherwise} \end{cases} \tag{7}$$

where λ_t is the trade-off parameter. Then, the adaptive attention loss is defined as the following:

$$\mathcal{L}_{att} = \rho_0 \left\|\mathrm{k}^* - \mathrm{k}_{gt}\right\|_1 + \rho_1 \left\|\mathrm{k}^*_{\lambda_1} - \mathrm{k}_{gt}\right\|_1 + \rho_2 \left\|\mathrm{k}^*_{\lambda_2} - \mathrm{k}_{gt}\right\|_1 \tag{8}$$

where λ_1 and λ_2 are the threshold value, ρ_0, ρ_1 and ρ_2 are the trade-off parameter to balance different loss terms. λ_1, λ_2, ρ_0, ρ_1 and ρ_2 are set as 0.01, 0.1, 0.01, 0.1 and 1 by experiences.

3. EXPERIMENTS

3.1. Datasets and Implementation Details

We collect 3450 high-quality 2K images from DIV2K [21] and Flickr2K [22] for training. Following previous blind SR methods [10], we synthesize the LR image by degradation with anisotropic Gaussian kernels. For scale factors $s \in \{2, 4\}$, the kernel sizes are set to 11×11 and 21×21. The range of kernel width σ_1 and σ_2 are set to $[0.2, 4.0]$, and the range of rotation angle is $[0, \pi]$.

During training, the degraded LR images are randomly cropped into 96×96 image patches and augmented with 90-degree rotations, random horizontal flips, and vertical. We also use Adam optimizer with mini-batch size 64 and 6×10^5 iterations. The initial learning rate is 1×10^{-2} and decays by a factor of 0.1 for every 2×10^5 iterations. We conduct all the experiments with the PyTorch framework on four NVIDIA Quadro P5000 graphics cards.

For evaluation, we use 4 widely used test benchmark datasets, including Set5 [17], Set14 [18], Bsd100 [19], and Urban100 [20]. Similar to other works, the SR performances are evaluated with PSNR and SSIM [23] on the Y channel in the YCbCr space.

3.2. Experiments on synthetic images

We compare the proposed KENet with bicubic interpolation and existing SOTA SR methods: RCAN [5], KernelGAN [11]+ZSSR [8], MZSR [9], DAN [16], and the upper bound model (non-blind USRNet [6] given ground truth kernels). RCAN is a SOTA bicubic PSNR-oriented method. MZSR is a zero-shot SR method for degradations with isotropic/anisotropic Gaussian kernels. KernelGAN+ZSSR combines a kernel estimation method and a non-blind SR method for blind SR. DAN is a blind SR method for degradations with isotropic/anisotropic Gaussian kernels. The test

Fig. 2. Visual comparisons of different SR methods on *img005* of the Urban100 dataset for scale factor $\times 2$.

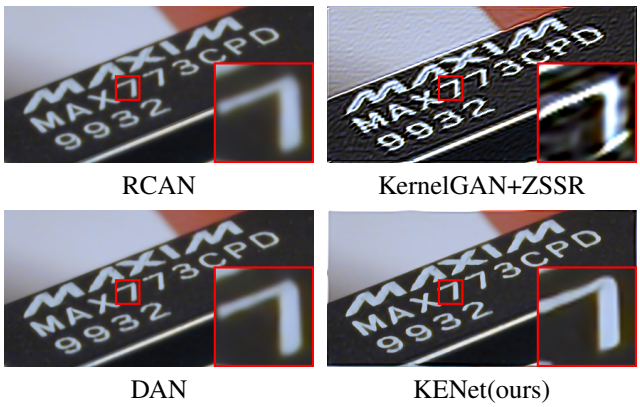

Fig. 3. Visual comparisons of different methods on super-resolving real image *chip* with scale factor $\times 4$.

LR images are randomly sampled with anisotropic Gaussian kernels $\sigma_1, \sigma_2 \in [0.2, 4.0]$. For a fair comparison, we use the publicly available source code to generate SR results.

The average PSNR/SSIM comparison of different SR methods is shown in Table 1. It can be observed that the proposed KENet achieves the best results than other SR methods. Specifically, our KENet outperforms other SR methods at least 3.36/0.89 dB PSNR for scale factors $\times 2$ and $\times 4$. Moreover, compared with the upper bound model, the proposed KENet only suffers 0.07 dB PSNR loss on the Bsd100 dataset for a scale factor of $\times 4$.

The qualitative evaluations are further shown in Fig. 2. We can observe from the zoom-in regions that the compared methods produce over-smoothing results or generate undesirable artifacts. In contrast, the proposed KENet produces a more pleasant result and is extremely similar to the ground-truth HR image.

3.3. Experiments on real images

We further compare the proposed KENet with RCAN, KernelGAN+ZSSR, DAN on a real image *chip* to demonstrate the SR performance. Since there is no ground-truth HR image for the real image *chip*, we only provide the visual comparisons. As shown in Fig. 3, KENet produces visually satisfactory results with clearer details and restores a higher contrast image.

Table 3. Quantitative results of different kernel estimation methods on the Set5 [17] dataset for scale factor $\times 2$.

Method	Kernel PSNR	Non-blind SR PSNR/SSIM
KernelGAN [11]+USRNet [6]	35.48	20.35/0.6237
DIPFKP [10]+USRNet [6]	37.46	28.27/0.8801
KENet+USRNet [6](ours)	**51.04**	**36.06/0.9503**
GT+USRNet [6](upper bound)	-	36.37/0.9514

3.4. Ablation study

3.4.1. Adaptive Attention Loss

To investigate the effectiveness of the adaptive attention loss $\mathcal{L}_{att}$, we retrain our model with $\mathcal{L}_1$ loss function for different scale factors. As shown in Table 2, KENet which is optimized with $\mathcal{L}_{att}$ outperforms the one with $\mathcal{L}_1$. Specifically, the proposed KENet with $\mathcal{L}_{att}$ obtains a gain of $1.99/11.25$ $(\times 2/4)$ dB PSNR for kernel estimation, and $0.07/0.09$ dB PSNR for blind SR, which demonstrates the effectiveness of $\mathcal{L}_{att}$.

3.4.2. Kernel Estimation Network

We further try to demonstrate the effectiveness of our KENet. We compare the proposed KENet with kernel estimation method KernelGAN and DIPFKP on the Set5 [17] dataset for scale factor $\times 2$. All the above kernel estimation methods are combined with the same non-blind SR model USRNet to achieve blind SR. A comparison of the estimated degradation kernels and blind SR performance is illustrated in Tabel 3. We can observe that our method achieves the highest performance both in kernel estimation and blind SR.

4. CONCLUSION

In this paper, we proposed a kernel estimation network named KENet. The proposed KENet estimates accurate degradation kernels and achieves SOTA blind SR performance when combined with non-blind SR methods. Experimental results on several image datasets have demonstrated the effectiveness of the proposed KENet.

5. REFERENCES

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[2] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[3] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1664–1673.

[4] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1132–1140.

[5] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, vol. 11211, pp. 294–310.

[6] Kai Zhang, Luc Van Gool, and Radu Timofte, "Deep unfolding network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3214–3223.

[7] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3262–3271.

[8] Assaf Shocher, Nadav Cohen, and Michal Irani, ""zero-shot" super-resolution using deep internal learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3118–3126.

[9] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3513–3522.

[10] Jingyun Liang, Kai Zhang, Shuhang Gu, Luc Van Gool, and Radu Timofte, "Flow-based kernel prior with application to blind super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10601–10610.

[11] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani, "Blind super-resolution kernel estimation using an internal-gan," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 284–293.

[12] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky, "Deep image prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9446–9454.

[13] Tomer Michaeli and Michal Irani, "Nonparametric blind super-resolution," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 945–952.

[14] Yossi Gandelsman, Assaf Shocher, and Michal Irani, ""double-dip": Unsupervised image decomposition via coupled deep-image-priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11026–11035.

[15] Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, and Wangmeng Zuo, "Neural blind deconvolution using deep priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3338–3347.

[16] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan, "Unfolding the alternating optimization for blind super resolution," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 5632–5643.

[17] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–10.

[18] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surfaces*, 2010, vol. 6920, pp. 711–730.

[19] Pablo Arbelaez, Michael Maire, Charless C. Fowlkes, and Jitendra Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, 2011.

[20] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5197–5206.

[21] Eirikur Agustsson and Radu Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1122–1131.

[22] Radu Timofte et al., "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1110–1121.

[23] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.