2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897356

ON ADVERSARIAL ROBUSTNESS OF DEEP IMAGE DEBLURRING

Kanchana Vaishnavi Gandikota *Paramanand Chandramouli* *Michael Moeller*

Department of Computer Science, University of Siegen

{kanchana.gandikota, paramanand.chandramouli, michael.moeller}@uni-siegen.de

ABSTRACT

Recent approaches employ deep learning-based solutions for the recovery of a sharp image from its blurry observation. This paper introduces adversarial attacks against deep learning-based image deblurring methods and evaluates the robustness of these neural networks to untargeted and targeted attacks. We demonstrate that imperceptible distortion can significantly degrade the performance of state-of-the-art deblurring networks, even producing drastically different content in the output, indicating the strong need to include adversarially robust training not only in classification but also for image recovery.

Index Terms— adversarial attack, image deblurring

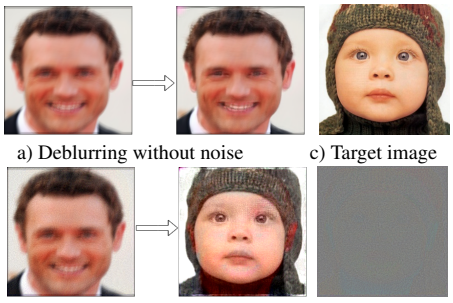

Fig. 1. Example targetted attack on DeblurGANv2[1].

1. INTRODUCTION

Image deblurring which aims at recovering sharp images from blurred inputs is an important and well studied research problem. Image blur occurs due to relative motion between cameras and objects in the scene during the exposure time, or due to suboptimal focal settings. The blur process can be mathematically represented as

$$y = B(x) + n, \tag{1}$$

where B is the blur operator producing blurry image y, x refers to the latent sharp image to be recovered, and n the additive noise. When the blur is uniform, the blur operation can be characterized using a convolution with blur kernel b

$$y = x * b + n, \tag{2}$$

Even for non-blind deblurring i.e with known blur operator, sharp image recovery is an ill-posed problem. When the blur operator is also unknown, it is referred to as blind deblurring, which is even more severely ill-posed, as multiple pairs of B and x can produce the same blurred observation y. Classical approaches to deblurring employ energy minimization with suitable priors, e.g. [2], and obtain the latent image x using iterative algorithms. Blind deblurring methods [3, 4] employ alternate minimization schemes to recover both x and B. Following the success of deep neural networks for computer vision applications, recently deep learning approaches have become the state of the art in image deblurring and other image restoration tasks [5, 1, 6, 7, 8]. While end to end trained deep networks can achieve impressive performances in many computer vision applications, these networks have been shown to be vulnerable to adversarial examples, wherein addition of carefully crafted imperceptible perturbation to the inputs can produce bizarre results [9, 10, 11]. Recent works [12, 13] have demonstrated adversarial attacks on medical image reconstruction and super-resolution.

In this work, we introduce adversarial attacks on image deblurring networks. We consider both blind and non-blind deblurring deep networks and evaluate their robustness to adversarial attacks. While most existing works focus on non-targeted attacks on image reconstruction, we investigate susceptibility of image deblurring networks to both targeted attacks and untargeted attacks. We demonstrate that state of the art deblurring networks are highly susceptible to adversarial perturbations. Fig. 1 shows an example targeted attack on DeblurGANv2 [1], a popular image deblurring network. A tiny additive perturbation to the input is sufficient to change the network output from the image of a man to that of a baby. While non-blind methods are relatively stable to such extreme changes in content, they can be susceptible to localized changes and untargeted attacks, indicating necessity for robust networks for image recovery.

2. RELATED WORK

Deep learning based image deblurring: Recently image restoration has witnessed a paradigm shift from classical ap-

proaches to using deep neural networks. We refer to [14] for a detailed survey and comparison of deep learning based image deblurring methods. Neural network approaches to blind deblurring typically learn to invert the blur operation directly using a trained neural network [1, 5] from large datasets of sharp and blurry image pairs to recover clean images. However, there are also methods which explicitly include the estimation of a blur operator [15, 16]. Even for non-blind deblurring, the knowledge of a blur operator has been successfully integrated into neural networks, by unrolling fixed steps of optimization algorithms with learned operators [6, 7, 17], or by using known deconvolution techniques in feature space [8] within the network. Recent works [18, 19] also take into account kernel uncertainty in non-blind deblurring. In addition to end to end trained networks for image recovery, neural networks are also used in iterative image recovery e.g. by using trained denoisers as proximal operators for plug and play reconstruction [20], or using trained generative priors [21].

Adversarial attacks on image reconstruction: While adversarial attacks on a neural network have been first introduced and extensively studied in the context of image classification [9, 10, 11], recent works [12, 22] have extended this to image reconstruction. While [12] investigates instabilities of MRI reconstruction networks by adding perturbations in the image domain, [22] consider adversarial perturbations in measurement domain and perform adversarial training using auxiliary network to generate adversarial examples. [23] perform adversarially attack to generate tiny features which cannot be recovered well by MRI reconstruction networks and propose adversarial training to improve the network's sensitivity to such features. [24] investigate adversarial robustness of different model based and model-inspired networks for CT and MRI reconstruction to untargeted attacks. They also investigate if networks can be attacked such that the resulting reconstruction causes a misclassification. In context of image restoration, [13] shows that several state of the art trained networks for image super resolution are susceptible to adversarial perturbations, however they consider only direct inversion models, with a focus on untargeted attacks. To the best of our knowledge, such attacks have not been shown for deblurring.

3. STABILITY OF IMAGE RECONSTRUCTION: ADVERSARIAL ATTACKS

Consider the problem of reconstructing x from the measurement model (2). An ideal reconstruction operator or a network $\mathcal{N}$ should have some notion of Lipschitz continuity such that small changes in the input produce only small bounded changes in the result.

$$\|\mathcal{N}(y_1) - \mathcal{N}(y_2)\| \leq K\|y_1 - y_2\|$$

Such error estimates exist (in terms of Bregman distances) in the case of convex energy minimization methods, (c.f. [25], Theorem 3.1 & the formula thereafter). While it is difficult to precisely characterize this notion for deep neural networks, prior works c.f. [26] attempt to approximate Lipschitz constant of neural networks. Moreover, instabilities could also arise from inaccurate estimate of the measurement operator, which can affect methods incorporating model knowledge into their architecture.

In this work, we investigate the robustness of image reconstruction networks to adversarial examples. Specifically, we craft adversarial examples by adding tiny norm bounded perturbations in the measurement domain. We consider only white box attacks where the parameters of the reconstruction network $\mathcal{N}$ are known to the attacker.

Untargeted Attacks: For a fixed image x, and a reconstruction network $\mathcal{N}$, the goal of an untargeted attack is to find an additive image perturbation that maximizes the reconstruction error subject to L_∞ constraints on the perturbation,

$$\delta_{adv} = \underset{\delta \in \mathrm{R}^m}{\operatorname{argmax}} \|\mathcal{N}(y+\delta) - \mathcal{N}(y)\|_2 \text{ s.t. } \|\delta\|_\infty \leq \epsilon. \quad (3)$$

Targeted Attacks: Here the goal is to find an additive image perturbation that produces a reconstruction close to a target image $\tilde{x}$ subject to L_∞ constraints on the perturbation,

$$\delta_{adv} = \underset{\delta \in \mathrm{R}^m}{\operatorname{argmin}} \|\mathcal{N}(y+\delta) - \tilde{x}\|_2 \text{ s.t. } \|\delta\|_\infty \leq \epsilon. \quad (4)$$

We solve the constrained optimization problems (3), (4) using the projected gradient descent (PGD) algorithm [11], with gradient updates using the Adam optimizer [27].

4. EXPERIMENTS

We use the following networks in our experiments: i) DeblurGANv2 [1] and ii) MPRNet [5], which are end to end trained networks for blind image deblurring, as well as iii) [6], a learned recurrent gradient descent network, and iv) [8], which performs Wiener deconvolution in the feature space of neural networks. The non-blind deblurring networks [6, 8] use the knowledge of blur operator during reconstruction. We use publicly available trained models of all these networks made available by the authors. For attacks on DeblurGANv2 [1], we choose the version using the inception backbone since it achieves the best results. In their experiments, [6] can unroll the recurrent gradient descent network for arbitrary number of steps during test time till a stopping criteria is satisfied. However, it becomes prohibitively complex to perform adversarial attack for high number of unrolled steps. In our experiments, we limit the number of unrolled steps to 10 for crafting adversarial inputs, but evaluate robustness to the same inputs using a network with 50 unrolled steps. We create a synthetic dataset of blurred images generated by convolving 8 uniform blur kernels of dataset in [28] with sharp images of different classes, 5 images CelebA-HQ dataset resized to 256×256, and 5 images from Berkeley segmentation dataset (BSDS300).

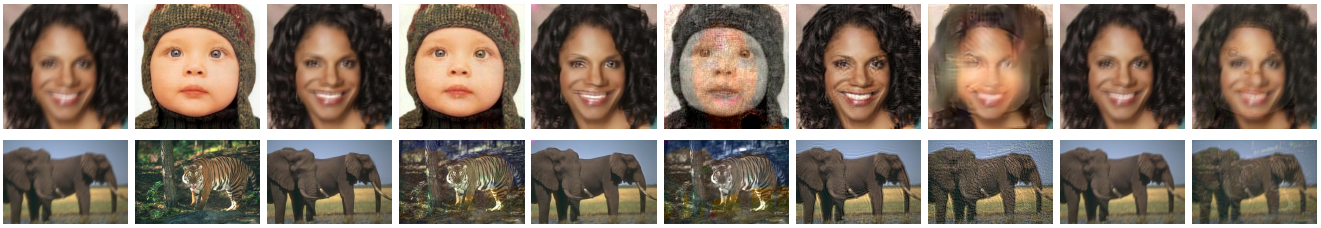

Fig. 2. Example targeted adversarial attacks on deblurring networks [5, 1, 8, 6].The columns 1 and 2 are blurred inputs and target images. For each network clean outputs (left), and network outputs with adversarial inputs (right) are depicted. Blurred images in rows 1 and 2 are generated using blur kernels '1' and '2' of size 19×19 and 17×17 in the dataset of [28] respectively.

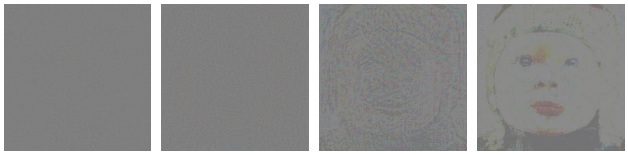

Fig. 3. Crafted adversarial perturbation for targeted attacks.

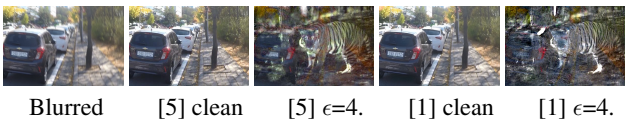

Fig. 4. Targeted attack with dynamically blurred input.

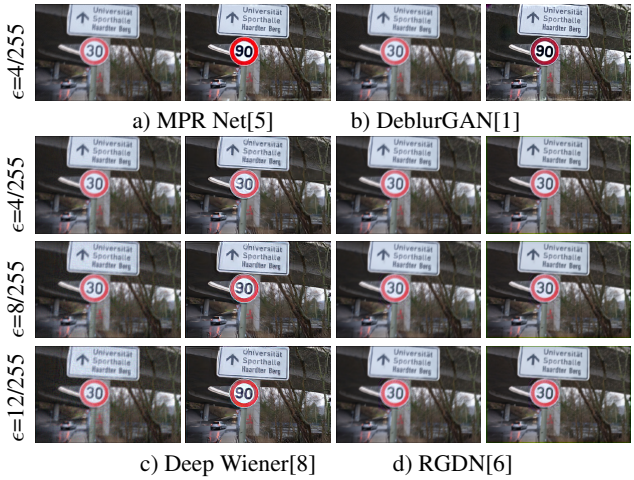

Fig. 5. Illustration of localized targeted attack. For each approach, the first column is the adversarial input and the second column is the network prediction.

These blur kernels have sizes ranging from 13×13 to 27×27. In our experiments, we use a fixed step size of $1e-3$ and use 250 PGD steps and 500 PGD steps to craft untargeted and targeted adversarial perturbations. We will make our data and code publicly available.

4.1. Targeted Attacks

In this section, we investigate the robustness of deblurring networks to targeted attacks that try to make the networks generate an image that is close to a target image. Fig. 2 illustrates example targeted attacks on the deblurring methods [5, 1, 8, 6]. Even though the blind deblurring methods [5, 1] are not trained using uniform blur models, they generate sharper images with less visible artifacts when the inputs are clean and the blur kernels are not too large. However, they are also most susceptible to targeted attacks, shockingly turning a woman into a baby or an elephant into a tiger with adversarial noise strength as low as 4/255.

In contrast, the non-blind methods are more robust and do not produce such extreme changes in the output, even for higher strengths of adversarial noise. The additive adversarial noise causing the targeted attacks is illustrated in Fig. 3. Adversarial perturbation for the non-blind methods show clear pattern of source images. However, the visual quality of the non-blind network outputs [8, 6] is lower even without adversarial noise. On our test data, the deep Wiener network [8] produces sharper results, but with visible ringing artifacts, and the recurrent gradient decent network [6] outputs still have a residual blur effect. Quantitative evaluation provided in Tab. 1 confirms the trend of higher susceptibility of the blind deblurring approaches to targeted attacks, showing higher similarity with the target image in terms of PSNR and normalized cross correlation (NCC) than with respect to the actual ground truth. While the blind dynamic deblurring approaches [5, 1] are not trained using uniform blur models, we find that similar adversarial vulnerabilities occur even with dynamically blurred images from the test sets of [5, 1], see Fig. 4.

We now investigate the susceptibility of networks to targeted attacks where target image is modified at a small localized region. Fig. 5 shows such a targeted attack on the deblurring networks, where the target image has the speed limit sign

Data	Net	Clean	Targeted attacks						Untargeted attacks		
			Similarity to source PSNR/NCC			Similarity to target PSNR/NCC			Similarity to source PSNR/NCC		
			$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 12$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 12$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 12$
Faces	[5]	26.81/0.976	9.77/0.409	9.75/ 0.408	9.75/0.408	26.80/ 0.986	26.94/0.987	26.94/0.987	11.29/ 0.645	8.88/0.498	8.582/0.465
	[1]	27.13/0.982	10.26/0.419	10.09/0.413	10.081/0.412	20.57/0.950	21.66/0.963	21.76/0.963	5.65/-0.063	5.36/-0.137	5.23/-0.175
	[8]	22.91/0.951	19.28/0.906	17.07/0.858	15.58/0.810	11.31/0.565	12.55/0.659	13.69/0.728	11.65/0.593	10.12/0.511	9.10/0.456
	[6]	26.98/0.981	24.55/0.961	23.32/0.954	22.06/0.934	10.19/0.452	10.57/0.496	10.95/0.537	24.13/0.966	22.67/0.953	21.43/0.939
BSD	[5]	24.41/ 0.944	12.52/0.155	12.30/0.130	12.27/0.126	23.15/0.899	24.14/0.919	24.22/0.921	10.59/0.487	9.04/0.398	8.12/0.319
	[1]	24.04/0.941	12.43/0.174	12.35/0.164	12.27/0.153	19.94/0.805	20.92/0.841	21.27/0.854	5.90/0.173	5.64/0.147	5.57/0.085
	[8]	21.44/0.443	21.08/0.882	19.43/0.844	18.08/0.792	13.11/0.139	14.10/0.223	15.00/0.311	14.15/0.607	11.43/0.478	10.15/0.416
	[6]	24.23/0.943	22.61/0.921	21.78/0.907	20.89/0.887	12.83/ 0.106	13.29/0.166	13.75/0.226	22.08/0.907	21.11/0.884	20.22/0.859

Table 1. Comparison of PSNR and normalized cross correlation (NCC) values with respect to source image for untargeted attacks, PSNR and NCC with respect to source and target images for targeted attacks to evaluate robustness.

Blurred [5] Clean [5] $\epsilon = 4$. [1] Clean [1] $\epsilon = 4$. [8] Clean [8] $\epsilon = 4$. [6] Clean [6]$\epsilon = 8$.

Fig. 6. Untargeted adversarial attack on deblurring networks. Blur kernel '6' of size 21×21 in the dataset of [28] is used.

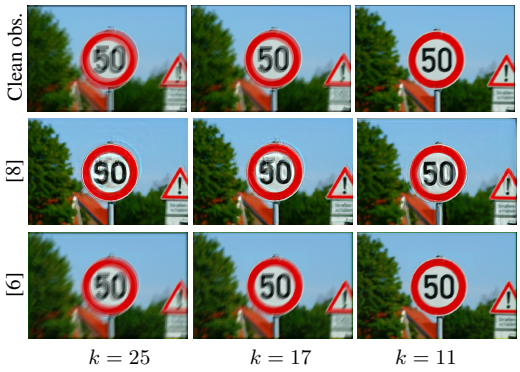

Fig. 7. Effect of kernel size on ease of targetted attack on non-blind deblurring networks with $\epsilon = 4/255$.

modified from '30' to '90'. The attack on blind deblurring networks is successful even at $\epsilon = 4/255$. Among the non-blind networks, attack on deep Wiener network [8] is clearly successful at $\epsilon = 12/255$, while the target features begin to manifest at even lower values of ϵ. The learned gradient descent approach [6] is most difficult to attack, barely producing target features even for $\epsilon = 12/255$. As the size of blur kernel becomes larger, deblurring becomes more ill-posed, which can influence the stability of the reconstruction. We investigate this effect by evaluating adversarial robustness of non-blind networks to targeted attacks by fixing adversarial noise level to 4/255, and varying the blur kernel size $\{25, 17, 11\}$. Here the target has a localized change in the speed limit sign from '50' to '90'. As the blur effect in the input reduces, we expect the networks to be more robust to attacks, which is confirmed by the results in Fig. 7. The robustness of deep Wiener deblurring [8] improves as the inputs are less and less blurred, and the learned gradient is least susceptible to attack.

4.2. Untargeted Attacks

In Tab. 1 and Fig. 6 we provide of effect of untargetted attacks on deblurring networks which increase the reconstruction loss. While the blind deblurring networks are highly susceptible to untargeted attacks, we find even the non-blind method of deep Wiener filtering also being unstable, even at low adversarial noise strengths.

In all our experiments, we find that the blind deblurring methods [1, 5] are most susceptible to adversarial perturbations. One reason is that blind deconvolution is inherently more ill-posed, making the reconstruction problem more unstable. Moreover, both the methods [1, 5] use only clean data during training, whereas the methods [6, 8] also add noise to blurry inputs during training. Recent work [24] shows addition of noise during training as an effective way to improve adversarial robustness of deep CT reconstruction. However, training with noise is not sufficient to guarantee adversarial robustness, as seen from the results of deep Wiener deconvolution [8], which is more prone to attacks than the learned gradient descent approach [6].

5. CONCLUSIONS

In this paper we introduced adversarial attacks on image deblurring networks and showed that state of the art deblurring methods can be highly susceptible to adversarial attacks. While the performance on clean data is important, it is critical to improve adversarial robustness of restoration approaches, by robust training or by developing more robust architectures.

6. REFERENCES

[1] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2019.

[2] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-laplacian priors," *Neurips*, vol. 22, 2009.

[3] M. Jin, S. Roth, and P. Favaro, "Normalized blind deconvolution," in *Proc. European Conference on Computer Vision*, 2018.

[4] L. Chen, F. Fang, T. Wang, and G. Zhang, "Blind image deblurring with local maximum gradient prior," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[5] S. W Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M. H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[6] D. Gong, Z. Zhang, Q. Shi, A. v. d. Hengel, C. Shen, and Y. Zhang, "Learning deep gradient descent optimization for image deconvolution," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 12, pp. 5468–5482, 2020.

[7] T. Eboli, J. Sun, and J. Ponce, "End-to-end interpretable learning of non-blind image deblurring," in *Proc. European Conference on Computer Vision*. Springer, 2020, pp. 314–331.

[8] J. Dong, S. Roth, and B. Schiele, "Deep wiener deconvolution: Wiener meets deep learning for image deblurring," *Neurips*, vol. 33, pp. 1048–1059, 2020.

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *Proc. ICLR*, 2014.

[10] I. J Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *Proc. ICLR*, 2015.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[12] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen, "On instabilities of deep learning in image reconstruction-does ai come at a cost?," *arXiv preprint arXiv:1902.05300*, 2019.

[13] J. H Choi, H. Zhang, J. H. Kim, C. J Hsieh, and J. S. Lee, "Evaluating robustness of deep image super-resolution against adversarial attacks," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2019.

[14] J. Koh, J. Lee, and S. Yoon, "Single-image deblurring with neural networks: A comparative survey," *Computer Vision and Image Understanding*, vol. 203, 2021.

[15] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf, "Learning to deblur," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1439–1451, 2015.

[16] A. Chakrabarti, "A neural approach to blind motion deblurring," in *Proc. European Conference on Computer Vision*, 2016, pp. 221–235.

[17] C. Bertocchi, E. Chouzenoux, M.C Corbineau, J. C Pesquet, and M. Prato, "Deep unfolding of a proximal interior point method for image restoration," *Inverse Problems*, vol. 36, no. 3, pp. 034005, 2020.

[18] S. Vasu, V. R. Maligireddy, and A. N. Rajagopalan, "Non-blind deblurring: Handling kernel uncertainty with cnns," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018.

[19] Y. Nan and H. Ji, "Deep learning for handling kernel/model uncertainty in image deconvolution," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[20] T. Meinhardt, M. Moller, C. Hazirbas, and D. Cremers, "Learning proximal operators: Using denoising networks for regularizing inverse imaging problems," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2017.

[21] M. Asim, F. Shamshad, and A. Ahmed, "Blind image deconvolution using deep generative priors," *IEEE Transactions on Computational Imaging*, vol. 6, 2020.

[22] A. Raj, Y. Bresler, and B. Li, "Improving robustness of deep-learning-based image reconstruction," in *Proc. ICML*, 2020, vol. 119 of *PMLR*, pp. 7932–7942.

[23] K. Cheng, F. Calivá, R. Shah, M. Han, S. Majumdar, and V. Pedoia, "Addressing the false negative problem of deep learning mri reconstruction models by adversarial attacks and robust training," in *Proc. 3rd Conference on Medical Imaging with Deep Learning*. 2020, PMLR.

[24] M. Genzel, J. Macdonald, and M. März, "Solving inverse problems with deep neural networks-robustness included," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[25] M. Burger, E. Resmerita, and L. He, "Error estimation for bregman iterations and inverse scale space methods in image restoration," *Computing*, vol. 81, no. 2, pp. 109–135, 2007.

[26] P. L Combettes and J. C Pesquet, "Lipschitz certificates for layered network structures driven by averaged activation operators," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 2, pp. 529–557, 2020.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[28] L. Sun, S. Cho, J. Wang, and J. Hays, "Edge-based blur kernel estimation using patch priors," in *Proc. IEEE International Conference on Computational Photography*, 2013.