

# IMPROVING SUPER RESOLUTION METHODS VIA INCREMENTAL RESIDUAL LEARNING

Muneeb Aadil, Rafia Rahim, Sibte ul Hussain

Reveal.ai Lab, National University of Computer and Emerging Sciences, Islamabad, Pakistan

{muneeb.aadil, rafia.rahim, sibteul.hussain}@nu.edu.pk

## ABSTRACT

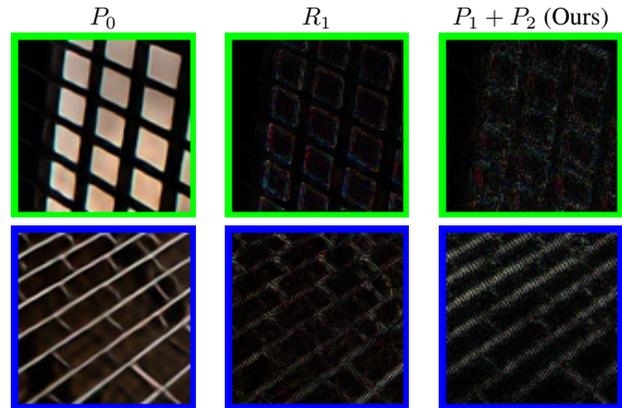
Recently, Convolutional Neural Networks (CNNs) have shown promising performance in super-resolution (SR). However, these methods operate primarily on Low Resolution (LR) inputs for memory efficiency but this limits, as we demonstrate, their ability to (i) model high frequency information; and (ii) smoothly translate from LR to High Resolution (HR) space. To this end, we propose a novel Incremental Residual Learning (IRL) framework to address these mentioned issues. In IRL, first we select a typical SR pre-trained network as a master branch. Next we sequentially train and add residual branches to the main branch, where each residual branch is learned to model accumulated residuals of all previous branches. We plug state of the art methods in IRL framework and demonstrate consistent performance improvement on public benchmark datasets to set a new state of the art for SR at only  $\approx 20\%$  increase in training time.<sup>1</sup>

**Index Terms**— Image Reconstruction, Convolutional Neural Networks, Residual Learning, Super Resolution

## 1. INTRODUCTION

Single Image Super-Resolution (SR) aims to generate a High Resolution (HR) image  $I^{SR}$  from a low resolution (LR) image  $I^{LR}$  such that it is similar to original HR image  $I^{HR}$ . SR has seen a lot of interest recently because it is: (i) inherently an ill-posed inverse problem; and (ii) an important low level vision problem having many applications.

In the wake of CNNs success on vision tasks, many researchers applied them to SR problems. One class of CNN based methods [1, 2, 3, 4, 5] take interpolated image as input and explicitly model residuals in image space. However, they have high memory and computational requirements since these networks take an interpolated HR version of an image as input. To tackle this issue, another class of methods [6, 7, 8, 9, 10] takes  $I^{LR}$  as input and apply convolutional operations primarily in LR space, followed by upsampling layers to produce  $I^{SR}$ . Although these networks do not suffer from high memory and computational cost, they have two design limitations. Firstly, these networks upsample the feature



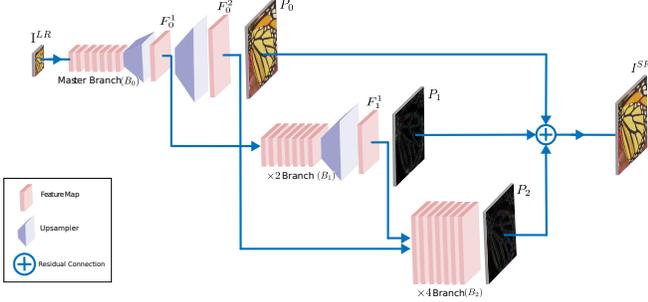
**Fig. 1:** Base network outputs  $I^{SR}$  (left) and their residuals (middle) *i.e.*  $I^{HR} - I^{SR}$  are learnt by IRL (right). These residuals are then added to main branch output for improved performance. Please refer to text for further details.

maps too swiftly and thus are unable to reliably learn the content gap between LR space and HR space. Secondly, these networks cannot explicitly learn residuals in image space because spatial dimensions of input and output do not match.

We hypothesize that: (i) there is a large content-gap between the LR feature maps and desired HR output image and this gap cannot be reliably learnt in a single transition step using a convolutional layer; (ii) instead of directly modeling HR image, one should explicitly model residuals – or High Frequency Information (HFI), in general – in image space to better model fine level details.

Thus, to this end, we propose the IRL framework: a new learning setup for SR task that improves the performance of existing networks. IRL uses an existing pre-trained network (*i.e.* master branch) and sequentially adds other networks (*i.e.* residual branches). Each added residual branch works on up-sampled feature maps, and is trained to learn the accumulated residuals of all the previous branches. At test time, the output of all branches are added to produce the final output. Our proposed architecture leads to consistent performance improvement of all the existing state of the art SR networks, adding only 20% extra training time and no extra memory overhead

<sup>1</sup>Code available at <https://bit.ly/2HipDsJ>



**Fig. 2:** Our proposed IRL framework for  $\times 4$  scale. Each branch represents an arbitrary image-to-image network. It is evolved from an existing architecture (master branch) by adding the residual branches ( $\times 2$  and  $\times 4$ ).

because of sequential training of the residuals.

## 2. RELATED WORK

We decompose our discussion in two separate topics and review them separately in the following sections.

### 2.1. CNNs for Super Resolution

SRCNN [1] first introduced CNNs for learning the SR task in end-to-end way and showed significant improvement over prior methods. VDSR [2] then tackled training difficulties encountered in SRCNN [1] by explicitly modeling residual learning in image space and proposed very deep 20 layers CNN. Similarly, DRCN [3] introduced parameters sharing in network using very deep recursive layers. Afterwards, DRRN [4] applied convolutional layers recursively to refine the image iteratively, and MemNet [5] employed memory block to take into account hierarchical information. However, since all these networks take as input an interpolated image, residual learning can be employed. But consequentially, these methods have high memory and computational cost.

To tackle this issue, recent approaches take input  $I^{LR}$  and apply convolutional operations primarily in LR space, followed by upsampling layers to produce  $I^{SR}$ . Since these methods apply majority of operations in the LR space, they do not suffer from high memory requirements and higher computational cost. For instance, SRResNet [6] adapts ResNet architecture [11] for SR and then employs ESPCNN [12] to up-sample efficiently. Lim *et al.* [7] improves SRResNet [6] and propose EDSR baseline (EDSRb) and EDSR. SRDenseNet [8] employs DenseNet architecture [13] coupled with skip-connections to explicitly model low level features. LapSRN [9] employs Laplacian pyramid style CNNs to iteratively process feature maps and upsample them to refine the earlier prediction. Then, RDN [10] merged residual block of EDSR [7] and dense skip connections of SRDenseNet [8] to form Residual Dense Block resulting in even better performance.

### 2.2. Objective Functions

Earlier networks majorly employ  $L_2$  loss because: (i) it directly optimizes PSNR, and (ii) it has nice optimization properties [14]. However, Zhao *et al.* [15] showed that  $L_1$  loss outperforms  $L_2$  in terms of PSNR. Qualitatively speaking,  $L_2$  recovers HFI *i.e.* edges better than  $L_1$  but it leaves splotchy artefacts on plain regions. On the other end,  $L_1$  loss removes splotchy artefacts at the cost of sharper edges recovery. Thus, following [15], majority of the networks employ  $L_1$  loss instead of  $L_2$  loss.

## 3. PROPOSED METHOD

In this section, we delineate our proposed framework and training methodology. Firstly, a SR network is selected as master branch  $B_0$  and is trained typically on  $L_0 = I^{HR}$ . When the training converges, its weights are frozen and another network *i.e.* residual branch  $B_1$  is added. This residual branch  $B_1$  takes the upsampled feature maps  $F_0^1$  of master branch  $B_0$  and is trained on the residuals  $R_1$  of master branch. This process is repeated until all scales of feature maps are processed – *c.f.* figure 2. Generally speaking, we can formalize the methodology as follows:

$$P_i = \begin{cases} B_i(I^{LR}) & \text{if } i = 0 \\ B_i([F_0^i, F_1^{i-1}, \dots, F_{i-1}^1]) & \text{if } i > 0 \end{cases} \quad (1)$$

Each branch  $B_i$  is trained on its respective label  $L_i$  which is formed as:

$$L_i = \begin{cases} I^{HR} & \text{if } i = 0 \\ R_i & \text{if } i > 0 \end{cases} \quad (2)$$

Where  $R_i = I^{HR} - \sum_{k=0}^{i-1} P_k$ . In the above equations, master branch and residual branches correspond to cases where  $i = 0$  and  $i > 0$ , respectively.

Finally, at test time, result  $I^{SR}$  is formed as:

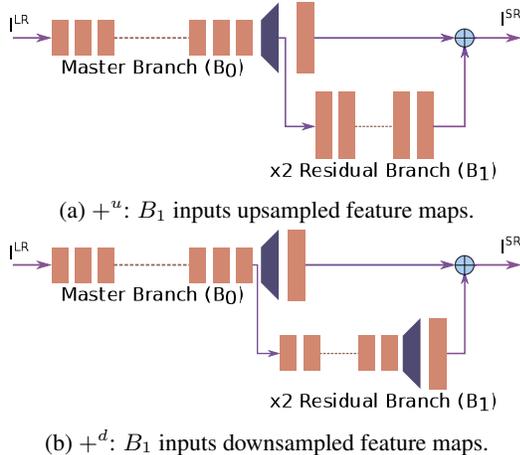
$$I^{SR} = \sum_{i=0}^n P_i \quad (3)$$

Where  $n$  is the number of branches and is dependent on SR scale. In our settings, we used a single branch for  $\times 2$ ,  $\times 3$  scales and 2 branches for  $\times 4$  scale.

Following sections explain in details the effects of each design choice of our architecture.

### 3.1. Upsampled Feature Maps

In our setup, each branch operates on higher dimensional feature maps than previous branch. We argue that this design choice leads to performance enhancement. To confirm this, we trained two variants, as shown in figure 3, and measured the performance difference between them. As shown in table 1,  $+^u$  versions lead to enhanced performance consistently



**Fig. 3:** Pictorial description of our  $+^u$  and  $+^d$  variants.

across all networks. Thus, for each branch, we used higher dimensional feature maps than previous branch. From now onwards,  $+/+$  versions take upsampled feature maps, unless specified otherwise.

Now since spatial dimensions are doubled after every branch, each successive branch’s parameters have to be halved also to keep the memory requirements constant. Therefore, we halved number of layers because it empirically showed the best results.

### 3.2. Residual Learning in Image Space

Although residual branches can be trained to learn  $I^{HR}$  directly, we instead train them on image residuals because it leads to faster convergence and slightly superior performance [2]. However, as opposed to [2] which used interpolated image for residuals computation, we used output of master branch  $P_0$  for  $B_1$ , and sum of  $P_0$  and  $P_1$  for  $B_2$ , since earlier branches are already trained in our scenario.

### 3.3. Extending Contemporary State of the Arts

To isolate the improvement IRL causes to existing state of the arts, we plug contemporary state of the art networks as our master branch and train only the residual branches from scratch. For uniform comparison, we keep training settings identical to the one used during the training of master branch.

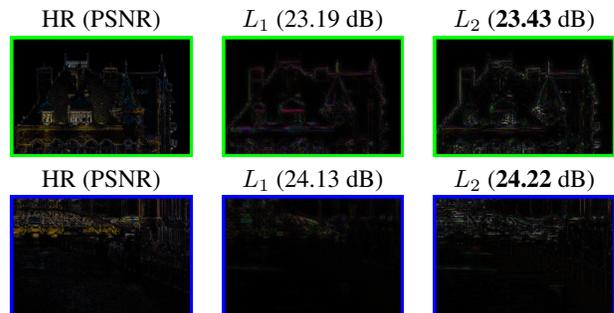
Table 1 shows the effects of adding IRL to the existing methods. IRL consistently improves the performance of all networks except the ones taking HR image as input. We argue this is because these networks already employ residual learning in image space and process upsampled feature maps.

### 3.4. Objective Functions

Existing methods have a trade-off between shaper edges and higher PSNR since it can be trained with only one loss func-

Network	Input	Original	IRL (Ours)		
			$+^d$	$+^u$	$++^u$
VDSR [2]	HR	30.58	N/A	30.58	30.58
DRRN [4]	HR	30.81	N/A	30.81	30.81
MemNet [5]	HR	30.86	N/A	30.86	30.86
LapSRN [9]	LR	30.67	30.69	<b>30.73</b>	<b>30.76</b>
SRResNet [6]	LR	29.80	29.81	<b>29.84</b>	<b>29.87</b>
EDSRb [7]	LR	29.83	29.84	<b>29.87</b>	<b>29.89</b>
EDSR [7]	LR	30.67	30.67	<b>30.69</b>	<b>30.71</b>
RDN [10]	LR	30.64	30.65	<b>30.67</b>	<b>30.69</b>

**Table 1:** Performance in terms of PSNR (dB) of  $+^d$  and  $++^u$  versions on validation set for  $\times 4$  scale. Red and blue denotes the best and second best performing methods respectively.



**Fig. 4:** Residuals  $R_1$  (left), learning capability of  $L_1$  loss (middle), and of  $L_2$  loss (right).  $L_2$  performs better than  $L_1$  for capturing details of residuals.

tion at a time due to having a single branch. In comparison since our framework has multiple branches, we can use different objective functions at the same time for modeling different information by each branch.

Specifically, different from the models trained in previous section, we trained the residual branches with  $L_2$  loss instead of  $L_1$  loss. Figure 4 shows the comparison:  $L_2$  loss not only recovers shaper edges, but it also performs better in terms of PSNR as opposed to previous works [7, 15]. We argue this is because previous works model HR image directly. On the other hand, we explicitly model HFI.

## 4. EXPERIMENTS AND RESULTS

We trained our network on DIV2K dataset [20] which contains 800 training images. We used first 790 images for training and the remaining 10 images for validation set. For testing, we used following benchmark datasets: Set5 [16], Set14 [17], B100 [18], and Urban100 [19].

For the training of residual branches, we used ADAM [21] optimizer with learning rate set to  $10^{-4}$  and  $\beta_1 = 0.9, \beta_2 = 0.99$ . We set input LR image patch size according to the one

Scale	Method	Set5 [16]	Set14 [17]	B100 [18]	Urban100 [19]
×2	LapSRN [9]	37.52 / 0.9573	33.08 / 0.9121	31.80 / 0.8924	30.41 / 0.9091
	LapSRN+ (Ours)	<b>37.55 / 0.9586</b>	<b>33.11 / 0.9122</b>	<b>31.82 / 0.8927</b>	<b>30.45 / 0.9093</b>
	EDSRb [7]	37.98 / 0.9604	33.56 / 0.9173	32.15 / 0.8994	31.97 / 0.9271
	EDSRb+ (Ours)	<b>38.01 / 0.9606</b>	<b>33.58 / 0.9174</b>	<b>32.18 / 0.8997</b>	<b>32.03 / 0.9276</b>
	EDSR [7]	38.11 / 0.9602	33.92 / 0.9195	32.32 / 0.9013	32.93 / 0.9351
	EDSR+ (Ours)	<b>38.14 / 0.9604</b>	<b>33.94 / 0.9196</b>	<b>32.34 / 0.9014</b>	<b>32.96 / 0.9354</b>
×3	RDN [10]	38.23 / 0.9613	34.00 / 0.9211	32.33 / 0.9016	32.87 / 0.9352
	RDN+ (Ours)	<b>38.27 / 0.9615</b>	<b>34.03 / 0.9214</b>	<b>32.36 / 0.9019</b>	<b>32.91 / 0.9355</b>
	EDSRb [7]	34.36 / 0.9267	30.28 / 0.8415	29.08 / 0.8053	28.14 / 0.8525
	EDSRb+ (Ours)	<b>34.41 / 0.9271</b>	<b>30.31 / 0.8417</b>	<b>29.11 / 0.8056</b>	<b>28.17 / 0.8528</b>
	EDSR [7]	34.65 / 0.9282	30.52 / 0.8462	29.25 / 0.8093	28.80 / 0.8653
	EDSR+ (Ours)	<b>34.68 / 0.9284</b>	<b>30.55 / 0.8465</b>	<b>29.28 / 0.8096</b>	<b>28.83 / 0.8655</b>
×4	RDN [10]	34.70 / 0.9293	30.56 / 0.8467	29.25 / 0.8094	28.79 / 0.8651
	RDN+ (Ours)	<b>34.73 / 0.9294</b>	<b>30.60 / 0.8469</b>	<b>29.27 / 0.8096</b>	<b>28.81 / 0.8654</b>
	LapSRN [9]	31.62 / 0.8869	28.12 / 0.7718	27.33 / 0.7286	25.28 / 0.7602
	LapSRN+ (Ours)	31.64 / 0.8867	28.16 / 0.7716	27.35 / 0.7289	25.33 / 0.7608
	LapSRN++ (Ours)	<b>31.67 / 0.8870</b>	<b>28.18 / 0.7718</b>	<b>27.37 / 0.7290</b>	<b>25.36 / 0.7612</b>
	SRResNet [6]	32.05 / 0.8910	28.53 / 0.7804	27.57 / 0.7354	26.07 / 0.7839
×4	SRResNet+ (Ours)	32.08 / 0.8913	28.55 / 0.7807	27.59 / 0.7356	26.13 / 0.7844
	SRResNet++ (Ours)	<b>32.10 / 0.8914</b>	<b>28.57 / 0.7808</b>	<b>27.62 / 0.7358</b>	<b>26.15 / 0.7845</b>
	EDSRb [7]	32.09 / 0.8926	28.56 / 0.7808	27.56 / 0.7359	26.03 / 0.7846
	EDSRb+ (Ours)	32.13 / 0.8932	28.59 / 0.7816	27.59 / 0.7362	26.07 / 0.7848
	EDSRb++ (Ours)	<b>32.15 / 0.8934</b>	<b>28.60 / 0.7817</b>	<b>27.60 / 0.7362</b>	<b>26.08 / 0.7849</b>
	EDSR [7]	32.46 / 0.8968	28.80 / 0.7876	27.72 / 0.7420	26.64 / 0.8033
	EDSR+ (Ours)	32.48 / 0.8970	28.83 / 0.7878	27.75 / 0.7423	26.66 / 0.8034
	EDSR++ (Ours)	<b>32.49 / 0.8971</b>	<b>28.84 / 0.7879</b>	<b>27.75 / 0.7424</b>	<b>26.69 / 0.8036</b>
	RDN [10]	32.46 / 0.8980	28.80 / 0.7868	27.71 / 0.7420	26.61 / 0.8027
	RDN+ (Ours)	32.48 / 0.8982	28.82 / 0.7871	27.74 / 0.7424	26.63 / 0.8030
	RDN++ (Ours)	<b>32.50 / 0.8982</b>	<b>28.84 / 0.7872</b>	<b>27.75 / 0.7424</b>	<b>26.64 / 0.8031</b>

**Table 2:** Quantitative Results (PSNR (dB) / SSIM) of adding IRL framework to the existing state of the art methods on different benchmarks datasets. Our proposed framework leads to consistent performance improvement for all scales.

used in the original master branch to ensure uniform comparison. For more details and qualitative results, please refer to our publicly available code. For quantitative comparison, we evaluated our approach in terms of PSNR and SSIM on Y channel against the following state of the art networks: LapSRN[9], SRResNet [6], EDSRb [7], EDSR [7], RDN [10] – *c.f.* Table 2. We skipped (i) SRDenseNet because of unavailability of pre-trained weights and public implementation of the method, and (ii) networks taking interpolated image as input because it already processes upsampled feature maps and learn residuals in image space.

## 5. CONCLUSIONS

Firstly, we argued that processing majorly in LR space is sub-optimal because (i) transition to HR space in single convolutional layer is unreliable, and (ii) it is not possible to learn residuals in image space which leads to performance loss.

Secondly, we empirically showed that for learning residuals,  $L_2$  loss performs better than  $L_1$  loss – unlike previous research – as it better models the HFI.

We addressed aforementioned issues by introducing residual branches which operate on upsampled feature maps, and learn residuals in image space through  $L_2$  loss. Our proposed framework leads to consistent improvement and sets a new state of the art for SR at marginal training time cost.

## 6. ACKNOWLEDGMENTS

We would like to acknowledge the computational grant from Higher Education Commission (HEC) of Pakistan.

## 7. REFERENCES

- [1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, “Image super-resolution using deep convo-

- lutional networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.
- [2] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, “Accurate image super-resolution using very deep convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.
- [3] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, “Deeply-recursive convolutional network for image super-resolution,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.
- [4] Ying Tai, Jian Yang, and Xiaoming Liu, “Image super-resolution via deep recursive residual network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [5] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu, “Memnet: A persistent memory network for image restoration,” in *Proceedings of International Conference on Computer Vision*, 2017.
- [6] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” *2017 IEEE CVPR*, pp. 105–114, 2017.
- [7] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, “Enhanced deep residual networks for single image super-resolution,” in *The IEEE CVPR Workshops*, July 2017.
- [8] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao, “Image super-resolution using dense skip connections,” *2017 IEEE ICCV*, pp. 4809–4817, 2017.
- [9] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, “Deep laplacian pyramid networks for fast and accurate superresolution,” in *IEEE CVPR*, 2017, vol. 2, p. 5.
- [10] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, “Residual dense network for image super-resolution,” in *CVPR*, 2018.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [12] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” *2016 IEEE CVPR*, pp. 1874–1883, 2016.
- [13] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, “Densely connected convolutional networks,” *2017 IEEE CVPR*, pp. 2261–2269, 2017.
- [14] Zhou Wang and A. C. Bovik, “Mean squared error: Love it or leave it? a new look at signal fidelity measures,” *IEEE Signal Processing Magazine*, vol. 26, pp. 98–117, 2009.
- [15] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz, “Loss functions for image restoration with neural networks,” *IEEE Transactions on Computational Imaging*, vol. 3, pp. 47–57, 2017.
- [16] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel, “Low-complexity single-image super-resolution based on nonnegative neighbor embedding,” 2012.
- [17] Roman Zeyde, Michael Elad, and Matan Protter, “On single image scale-up using sparse-representations,” in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.
- [18] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. IEEE, 2001, vol. 2, pp. 416–423.
- [19] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, “Single image super-resolution from transformed self-exemplars,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5197–5206.
- [20] Eirikur Agustsson and Radu Timofte, “Ntire 2017 challenge on single image super-resolution: Dataset and study,” in *The IEEE CVPR Workshops*, July 2017.
- [21] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.