LADDER PYRAMID NETWORKS FOR SINGLE IMAGE SUPER-RESOLUTION

Zitao Mo[1,2] *Xiangyu He*[1,2] *Gang Li*[1,2] *Jian Cheng*[1,2,3]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Center for Excellence in Brain Science and Intelligence Technology, CAS, Beijing, China
mozitao2017@ia.ac.cn, {xiangyu.he, gang.li, jcheng}@nlpr.ia.ac.cn

ABSTRACT

Benefiting from the powerful representation capability of convolutional neural networks, the performance of single image super-resolution (SISR) has been substantially improved in recent years. However, many current CNN-based methods are computation-intensive because of large-size intermediate feature maps and inefficient convolutions. To resolve these problems, we propose Ladder Pyramid Network (LPN) for single image super-resolution. Firstly, we use strided convolution to reduce the size of the intermediate feature maps and thus reducing computation burden. In order to better balance the effectiveness and efficiency, we propose Ladder Pyramid Module to gradually fuse hierarchical features to enhance performance. Secondly, lightweight convolution block similar to Inverted Residual Module of Mobilenet-v2 was introduced into SISR, with which we build the network backbone and ladder feature pyramid. Experimental results demonstrate that the proposed Ladder Pyramid Network can achieve comparable or better performance than previous lightweight networks while reducing the amount of computation.

Index Terms— Ladder Pyramid Network, Lightweight Convolution, Super-Resolution

1. INTRODUCTION

Single image super-resolution (SISR) aims to reconstruct high-resolution images from a single low-resolution sample. SISR has been an active research topic for its widely applications on facial image improvement [1], satellite and aerial imaging [2], etc. However, this problem is non-trival, since different high-resolution images may generate the same low-resolution sample [3]. Despite of its ill-posed essence for many-to-one mapping, rapid developments of deep convolutional neural networks push forward progress in this field.

Ever since Dong et al. [4] first applied convolutional neural network to SISR, deeper and more powerful convolutional neural networks have been developed to enhance performance. Although significant improvements have been witnessed, most of the current methods are computation-intensive, which poses a challenge to the real-world applications. There are two main reasons for the large amount of computation in previous work. Firstly, for most of the CNN-based methods, SISR is inherently computation-intensive because it relies on the convolutions of large-size intermediate feature maps. In order to maintain detailed information of original images, many previous works tend not to downsample the input image [5, 6, 7, 8] or even upsample the image [9] in the early layers, imposing a huge computation burden. Secondly, the application of lightweight convolution (e.g. depthwise separable convolution) in SISR is largely overlooked, although they have presented high efficiency in other areas such as classification, detection and segmentation [10]. Most of the prior arts design deep convolutional neural networks based on normal convolution, which may not be a good choice for efficient computation.

In this paper, we get started with the above two issues that give rise to the huge computation burden and design efficient convolutional neural networks for SISR. Firstly, unlike some previous work, we use strided convolution (stride > 1) to reduce the resolution of intermediate feature maps and thus can decrease the amount of computation. For low-level tasks, downsampling the intermediate feature maps may lead to loss of details and performance degradation. In order to compensate for the loss of information, we propose Ladder Pyramid Network, which can gradually merge high-level information and low-level details to enhance performance. Secondly, we introduce lightweight Inverted Residual Module [10] to SISR, and modify it by removing the batch normalization layer and retaining only one activation function to accommodate to this task. The module is used in both the backbone and the pyramid module to achieve a good balance between performance and efficiency. The contributions are summarized below:

- We propose Ladder Pyramid Network for SISR. We use strided convolution to reduce computation and develop Ladder Pyramid Module to incorporate hierarchical information to enhance performance.

This work was supported in part by National Natural Science Foundation of China (No.61972396, 61876182, 61906193), the Strategic Priority Research Program of Chinese Academy of Science (No.XDB32050200), the Advance Research Program (31511130301).

- We introduce and modify Inverted Residual Module to construct the backbone and Ladder Pyramid Module, which can further reduce computation burden. Experimental results demonstrate that the proposed Ladder Pyramid Network can achieve comparable or better performance than prior arts with less multiply-accumulate operations.

2. RELATED WORK

Since AlexNet [11] made a breakthrough in large-scale image classification competitions, other areas of computer vision have also been inspired to start developing methods based on convolutional neural networks, including single-image super-resolution. Dong et al. [4] first introduces convolutional neural networks for SISR and achieved superior performance than traditional methods. Kim et al. [9] then goes further with very deep convolutional networks, however, they firstly interpolate the original images to the desired resolution, which brings about a huge amount of operations. To reduce computational complexity, FSRCNN [5] and ESPCN [6] propose to upsample images at the end of the networks. In order to fully harness the potential of deep architectures, EDSR [7] modifies SRResNet [12] by removing unnecessary module and carefully engineering the architecture, which brings about enhanced performance. CARN [3] introduces multiple shortcut connections to incorporate the features from multiple layers, which achieve better trade-off between performance and computation overhead. OISR [8] introduces an ODE-inspired scheme to develop convolutional neural networks for SISR, however, they present superior performance than CARN [3] at a higher number of multiply-accumulate operations.

Another topic related to this paper is feature pyramid network (FPN) [13]. Since traditional image pyramid is inefficient, feature pyramid network is developed for object detection to incorporate inherent multi-scale features of convolutional neural networks, which can improve performance with small extra cost. Then Liu et al. [14] improve FPN and propose PANet to achieve better performance on instance segmentation. Inspired by traditional laplacian pyramid, Lai et al. [15] propose Laplacian Pyramid Super-Resolution Network (LapSRN), however, the performance of LapSRN is worse than other methods due to the naive architectures. Kirillov et al. [16] introduce FPN to panoptic segmentation, the simple strategy achieves state-of-the-art performance for this dense-pixel prediction task. For another pixel-level task image deblurring, Kupyn et al. propose DeblurGAN-v2 to explore the application of FPN on image restoration [17].

3. LADDER PYRAMID NETWORK

In this section, we develop Ladder Pyramid Network for single image super-resolution. We use strided convolution to reduce resolution of intermediate feature maps, then propose Ladder Pyramid Module to incorporate information from different layers and introduce lightweight convolution module to further reduce multiply-accumulate operations. The overall architecture is based on Convolution-PixelShuffle framework similar to EDSR [7], except that we replace the body with our Ladder Pyramid Network, as depicted in Fig. 1.

Ladder Pyramid Module: Since the intensive computation of SISR is largely due to the high-resolution features, we use strided convolution to obtain downsampled intermediate feature maps. However, low-resolution features may suffer from loss of detailed information, we resolve it by building Ladder Pyramid Module, which can incorporate multi-level information. Unlike FPN, the Ladder Pyramid Module consists of multiple top-down pathways (α, β, γ) to gradually incorporate hierarchical information. Intermediate features (C_1, C_2, C_3, C_4) with scale of 1, $1/2$, $1/4$ and $1/8$ encoded by the backbone are fed into the Ladder Pyramid Module. In each pathway, we perform convolution and $2\times$ upsampling for higher-level features and add them to lower-level feature maps. The conv-upsample-add strategy is repeated for feature maps with different scales, which can be formulated as

$$C_k^{\alpha} = U(C_{k+1}^{\alpha}) + \phi(C_k),\ k = 1, 2, 3, \tag{1}$$

$$C_k^{\beta} = U(C_{k+1}^{\beta}) + \phi(C_k^{\alpha}),\ k = 1, 2, \tag{2}$$

$$C_k^{\gamma} = U(C_{k+1}^{\gamma}) + \phi(C_k^{\beta}),\ k = 1, \tag{3}$$

in which U denotes $\times 2$ upsampling and ϕ denotes a convolution module. Finally, the output of each top-down pathway $(C_1^{\alpha}, C_1^{\beta}, C_1^{\gamma})$ together with high-resolution features from the backbone (C_1) are concatenated together and fed into another two convolutional layers.

In order to understand the effects of Ladder Pyramid Module, we visualize the feature maps of conv-upsample-add module (Fig. 1). As the figure depicts, high level features focus on semantic information (the head and body of a bird) but lose some detailed information, while the feature maps from lower level present more local details (contour and texture). The Ladder Pyramid structure can compensate for the loss of detailed information.

Modified Inverted Residual Module: To further reduce computation, we introduce lightweight Inverted Residual Module to SISR and plug it into both the backbone and ladder pyramid module. However, modules designed for other visual tasks are not necessarily the best choice for single image super-resolution. We have made some modifications to accommodate to SISR (Fig. 2). Firstly, as suggested by EDSR [3, 7], batch normalization layers are removed from the module (Fig. 2 (a)), which will also accelerate the training. Secondly, we use ReLU as activation function instead of ReLU6 (Fig. 2 (b)), which is found to improve performance (Section 4). We also find that not all activations are necessary, the performance can be slightly enhanced by retaining only one activation (Fig. 2 (c)).

Detailed Architecture: The backbone of the network

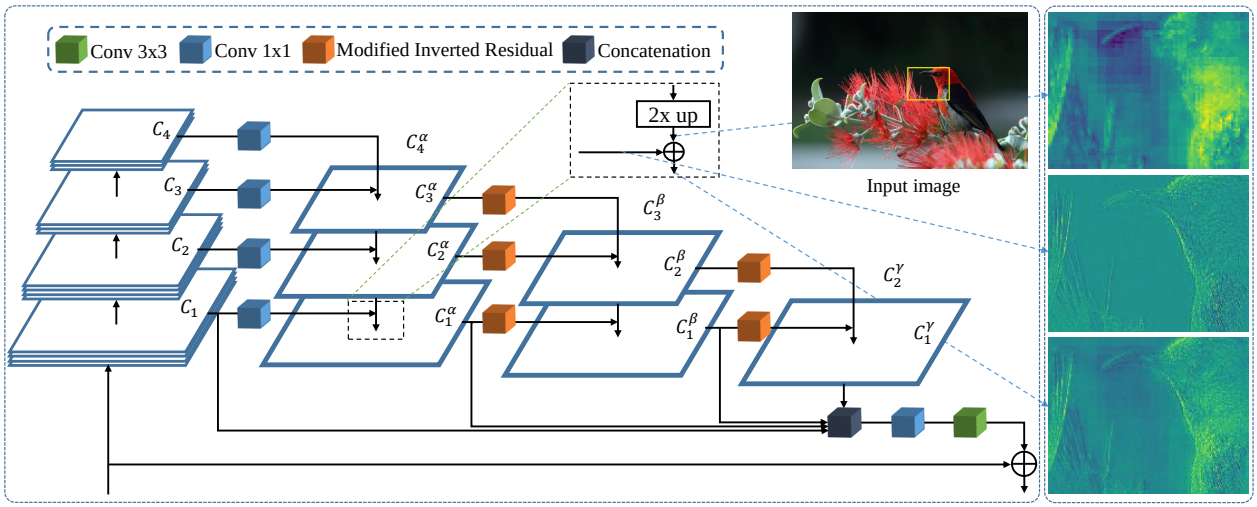

Fig. 1. **Left:** Ladder Pyramid Network (LPN). Unlike FPN, we adopt three top-down pathways to gradually incorporate hierarchical features. From left to right we denote them as pathway-α (Eq. (1)), pathway-β (Eq. (2)) and pathway-γ (Eq. (3)), and in Section 4 we will explore their impact on performance. **Right:** Extracted features from the network. Top: $Upsample(C_2^\alpha)$. Middle: $Conv(C_1)$. Bottom: C_1^α. We can notice that high-level layer extracts quite coarse feature and focus on certain part of the bird, while low-level feature maintains more fine-grain information and compensate for the detail loss.

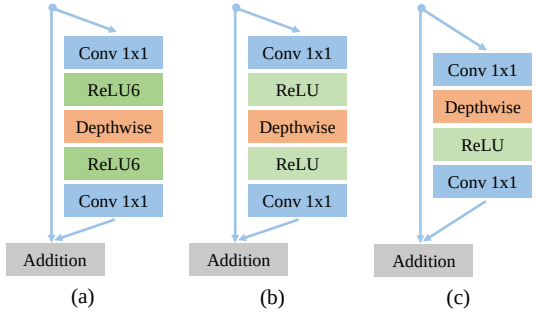

Fig. 2. Comparisons of three convolution modules. Module (c) presents better performance than the other two architectures (see in Section 4).

consists of four stages, the number of layers in each stages are 4, 4, 3, 3, and the number of output channels of each layers are set to be 64, 96, 128, 160, respectively. The expansion ratio of modified Inverted Residual Module is 4, except that we use a larger expansion ratio of 12 in the first block of each stage. We also add an additional shortcut from the second to the last block in each stage, which encourages quick information propagation from lower to higher layers. As for Ladder Pyramid Module, we adopt an expansion ratio of 6. The kernel size of depthwise convolution is 3 while we use kernel size of 5 in the backbone.

4. EXPERIMENTS

4.1. Experimental Settings

Similar to previous works [7, 8], we train our networks with 1st-800th images in DIV2K and validate on 801st-810th images. All the input images are augmented with the same settings as [7, 8]. The models are trained with ADAM optimizer by setting $\alpha_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ for 800 epochs with minibatch of 16. We set the initial learning rate to 4×10^{-4} and halved every 250 epochs. We use the RGB input patches of size 48×48 from the low-resolution image. We report peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) of Y channel (YCbCr space) on four benchmark datasets Set5, Set14, B100, and Urban100.

4.2. Results on Benchmark Datasets

We compare our results with MemNet [18], SelNet [19], CARN [3] and OISR [8], whose computations are in a similar number of magnitude. We also compare with FALSR [20] based on Neural Architecture Search. As shown in Table 1, when the upscaling factor is 2 or 3, LPN outperforms MemNet, SelNet, CARN and FALSR in most cases with fewer multiply-accumulate operations. The performance are also comparable with OISR-RK2-s even though the computation of LPN is much smaller. When the upscaling factor is 4, LPN performs better than SelNet and MemNet, but falls behind CARN and OISR-RK2-s in some cases. This may be due to the use of strided convolution, some detailed information is missing even though we use a pyramid network. Nonetheless, LPN still achieves comparable performance with CARN.

4.3. Model Analysis

In this part, we explore the influence of convolution modules (Fig. 2), top-down pathways (Fig. 1) and channel numbers.

Effects of convolution module: We compare three architectures shown in Fig. 2. Experimental results on Table 2 show that module (c) achieves better performance than the

Table 1. Quantitative results (PSNR(dB) / SSIM) of our model (LPN) compared with other lightweight networks on benchmark datasets. "MAC" denotes the number of multiply-accumulate operations calculatd by assuming that the output images are 720×1280. The **bold** denotes the best-performing models (higher PSNR/SSIM or smaller MAC) while the underline denotes the second. Note that FALSR only reports ×2 results.

Method	Scale	MAC	Set5		Set14		B100		Urban100	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MemNet [18]	×2	623.9G	37.78	0.9597	33.28	0.9143	32.08	0.8978	31.31	0.9195
SelNet [19]	×2	225.7G	37.89	0.9598	33.61	0.9160	32.08	0.8984	–	–
CARN [3]	×2	222.8G	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256
OISR-RK2-s [8]	×2	316.2G	**37.98**	0.9604	33.58	0.9172	**32.18**	**0.8996**	**32.09**	0.9281
FALSR-A [20]	×2	234.7G	37.82	0.9595	33.55	0.9168	32.12	0.8987	31.93	0.9256
LPN	×2	**184.5G**	37.97	**0.9606**	**33.63**	**0.9186**	32.14	0.8994	32.06	**0.9287**
MemNet [18]	×3	623.9G	34.09	0.9248	30.00	0.8385	28.96	0.8001	27.56	0.8376
SelNet [19]	×3	120.0G	34.27	0.9257	30.30	0.8399	28.97	0.8025	–	–
CARN [3]	×3	118.8G	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493
OISR-RK2-s [8]	×3	160.1G	**34.43**	**0.9273**	30.33	0.8420	**29.10**	**0.8053**	**28.20**	**0.8534**
LPN	×3	**101.7G**	34.31	0.9259	**30.34**	**0.8429**	29.07	0.8050	28.09	0.8514
MemNet [18]	×4	623.9G	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630
SelNet [19]	×4	83.1G	32.00	0.8931	28.49	0.7783	27.44	0.7325	–	–
CARN [3]	×4	90.9G	32.13	0.8937	28.60	0.7806	**27.58**	0.7349	26.07	0.7837
OISR-RK2-s [8]	×4	114.2G	**32.21**	**0.8950**	**28.63**	**0.7822**	**27.58**	**0.7364**	**26.14**	**0.7874**
LPN	×4	**81.3G**	32.16	0.8928	28.60	0.7817	27.55	**0.7364**	25.95	0.7822

Table 2. Effects of convolution module (Fig. 2) and top-down pathways (Fig. 1). We report PSNR of scale ×2 on Set14.

Models	LPN-(a)	LPN-(b)	LPN-(c)(LPN)
MAC	184.5G	184.5G	184.5G
Params	2.64M	2.64M	2.64M
Set14	33.56	33.62	**33.63**
Models	LPN-α	LPN-β	CARN
MAC	151.6G	168.4G	222.2G
Params	2.37M	2.53M	1.59M
Set14	33.52	33.57	33.52

other two structures. Therefore, it is used as the basic block of our models.

Effects of top-down pathway: We construct models with only one top-down pathway (pathway-α, **LPN-α**) and two top-down pathways (pathway-α and pathway-β, **LPN-β**), and compare them with original Ladder Pyramid Network (**LPN**), as presented in Table 2. We can see that Ladder Pyramid Network with all three top-down pathways performs better than another two models with small extra computation overhead, which validates the effectiveness of the ladder architecture. We also find that even LPN-α performs the same as CARN with much smaller operations.

Scalability of LPN: Finally, we explore the scalability of the proposed models by setting the channels to be 1/2 (**LPN-0.5**) and 3/4 (**LPN-0.75**) of the origin models. We compare LPN family with other CNN-based benchmark algorithms in terms of numbers of operations (MAC) and PSNR, as shown in Fig. 3. LPN presents superior performance when the upscale factor is ×2. Specifically, it achieves better performance-computation trade-off than CARN.

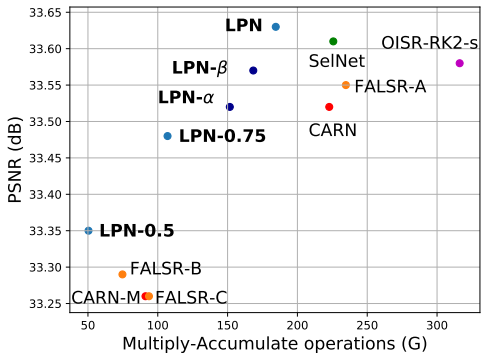

Fig. 3. The proposed models (the **bold**) and comparisions with other methods on Set14. The scale factor is ×2.

5. CONCLUSION

In this paper, we propose Ladder Pyramid Network for single image super-resolution. Inverted Residual Module is introduced and modified to accommodate to SISR, with which we can build a model with smaller computation burden. Experimental results on 4 benchmark datasets verify the efficiency of our models. In the future, we will explore the combination of the proposed network and attention mechanism or knowledge distillation to build a stronger model. Applications of Ladder Pyramid Network in other task such as semantic segmentation are also expected to be explored.

6. REFERENCES

[1] Bo Li, Hong Chang, Shiguang Shan, and Xilin Chen, "Low-resolution face recognition via coupled locality preserving mappings," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 20–23, 2010.

[2] Yun Zhang, "Problems in the fusion of commercial high-resolution satellite as well as LANDSAT 7 images and initial solutions," 2002.

[3] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 256–272.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*, 2014, pp. 184–199.

[5] Chao Dong, Chen Change Loy, and Xiaoou Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*, 2016, pp. 391–407.

[6] 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," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1874–1883.

[7] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1132–1140.

[8] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng, "ODE-Inspired network design for single image super-resolution," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1732–1741.

[9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1646–1654.

[10] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.

[12] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.

[13] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[14] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia, "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.

[15] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5835–5843.

[16] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollar, "Panoptic feature pyramid networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6399–6408.

[17] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang, "DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8878–8887.

[18] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu, "Memnet: A persistent memory network for image restoration," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4549–4557.

[19] Jae-Seok Choi and Munchurl Kim, "A deep convolutional neural network with selection units for super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 1150–1156.

[20] Xiangxiang Chu, Bo Zhang, Hailong Ma, Ruijun Xu, Jixiang Li, and Qingyuan Li, "Fast, accurate and lightweight super-resolution with neural architecture search.," *arXiv preprint arXiv:1901.07261*, 2019.