

ENHANCING DIGITAL ZOOM IN MOBILE PHONE CAMERAS BY LOW COMPLEXITY SUPER-RESOLUTION

Farzad Toutounchi, Ebroul Izquierdo

Multimedia and Vision Research Group, Queen Mary University of London

{f.toutounchi, e.izquierdo}@qmul.ac.uk

ABSTRACT

In this paper, we present an enhancement method for improvement of digital zoom feature in mobile phone cameras. The enhancement approach is based on the state-of-the-art deep learning-based super-resolution. We introduce an input super-pixel pooling layer in the enhancement framework, that down-scales the input data without losing information, and results in a 96% complexity reduction of the enhancement framework, when compared with the baseline approach. The presented model is also capable of being trained simultaneously for several scaling factors and different levels of enhancement, without affecting the complexity of the approach.

Index Terms— Computational Photography, Super-Resolution, Deep Learning

1. INTRODUCTION

The recent advances in smart phones technology and the improvements of the cameras embedded in the mobile devices allow users to take high quality images with very high spatial resolutions. A camera lens with 16 million pixels, capable of capturing still images with a spatial resolution of 5312×2988 , is a very typical case in the existing smart phones in the market, that can provide very high quality digital photography.

An important and appealing feature for users in digital cameras is the zooming functionality that can enable them to get close-up views of the scenes and provide a photographic degree of freedom. Zooming can be categorized into two different classes of optical zoom and digital zoom based on the adopted technology. Optical zoom is performed by changing the focal length of a zoom lens, and is widely available in professional cameras and can promise a consistent quality of image from different angles. Digital zoom, on the other hand, handles the zooming operation by selecting (cropping) a portion of the pixels in the camera lens and performing an interpolation to reach the desired image size.

Digital zoom is the method of choice in smart phones due to the existing limitations in the hardware, and as experienced by any naïve user, it can not guarantee a very high quality imaging when compared with the optical zoom. This shortcoming triggered our research in computational solutions for

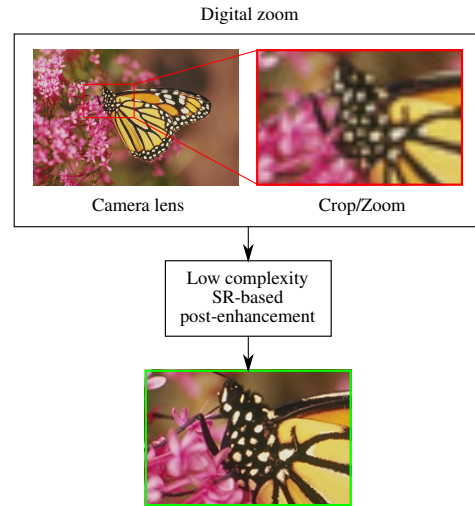


Fig. 1. General work-flow of digital zoom and the proposed quality enhancement process.

enhancing the quality of the digital zoom in mobile cameras.

Inspired by the recent developments in application of deep learning for still image Super-Resolution (SR), we propose a low complexity post-processing step in the computational photography work-flow of the mobile phones, as depicted in Fig. 1, that can improve the digital zoom quality significantly.

We introduce a super-pixel pooling layer applied to a state-of-the-art SR approach for splitting the image into several smaller ones, hence reducing the complexity of convolutional operations on the input data. The enhancement is then performed by the SR network, and the image is reconstructed to its original size using a sub-pixel up-scaling layer [1]. The presented low complexity model can provide the same quality image enhancement in comparison with the baseline approach, and unlike the existing low complexity models, can be trained and applied for any scaling factors.

The rest of this paper is organized as follows. Section 2 summarizes the existing deep learning-based approaches in still image SR. Section 3 describes the proposed SR method for digital zoom enhancement in details. Section 4 reports on the performance of the approach in terms of quality and complexity, followed by conclusions in Section 5.

2. DEEP LEARNING IMAGE SUPER-RESOLUTION

Dong et al. [2, 3] presented the first concrete deep learning-based SR approach for still images, which exploited deep Convolutional Neural Networks (CNN), capable of up-sampling still images with high visual quality and reasonable complexity. They improved their work further in [4] by introduction of transposed convolutional layers to their framework for up-sampling from low to high resolution. Including the transposed convolution as the last layer enables the network to perform the operations in the low resolution, hence reducing the complexity significantly, while the image is up-sampled in the last layer without needing the bi-cubic interpolation pre-processing step. Shi et al. [1] used a similar concept as [4] for reducing the complexity of the framework, except they introduced a sub-pixel up-scaling layer instead of the transposed convolution layer for the up-sampling part, with which they achieved an efficient performance. Both [1] and [4], although very effective in reducing the SR complexity, share a critical deficiency, which is the inconsistency of the network structure for different scaling factors. These models require different structures for each scaling factor, and that leads to dependence of the complexity to the scaling factor, as well as the input image resolution.

Other major contributions in still image SR using deep learning have been made by Kim et al. [5, 6]. In [5], the concept of residual learning was introduced to deep SR models, which led to quicker convergence of the network and high visual quality reconstruction. In [6], a deep and recursive architecture was employed, that achieved very good performance in terms of image quality, although expensive in terms of computation cost. In other deep learning-related efforts, Cui et al. [7] employed a cascade of multiple collaborative local auto-encoders that up-samples the low-resolution image gradually and layer by layer. Wang et al. [8] introduced a deep joint SR model to exploit both external and self-similarities for reconstruction, in which a stacked de-noising convolutional auto-encoder was first trained on external training data, then it was fine-tuned with multi-scale self-examples from the input data.

Another interesting endeavor in devising deep learning models for image SR was done by Mao et al. [9]. They introduced a very deep hourglass-shaped CNN that included multiple convolution and transposed convolution layers and was originally designed for image de-noising. However, the model can be applied for SR and image up-sampling as well, and is reported in the literature as one of the best SR models for high quality image up-sampling.

3. IMAGE SUPER-RESOLUTION WITH SUPER-PIXEL POOLING LAYER

Digital zoom is essentially performed by spatial up-sampling of a cropped image in mobile phone cameras. Hence SR can be exploited as a tool for enhancing this feature in smart phones. The application of SR, thus, can be considered as an enhancement stage on the already zoomed-in and up-sampled image. Given that most of the deep learning-based SR approaches perform a bi-cubic interpolation as a pre-processing step, the built-in interpolation process in the cameras can also be considered as a pre-processing step prior to the SR-based enhancement stage. Consequently, we would require a CNN that can map a low quality image to a higher quality version with the same resolution.

As mentioned earlier, today's mobile cameras provide very high resolution images, and performing deep learning-based SR for such resolutions is computationally expensive, and requires major processing power. This means the existing state-of-the-art SR approaches, although very promising in terms of quality, cannot be integrated in smart phones, as they can function very slowly on CPUs. This inspired us to propose a network architecture that can still perform reasonably fast on CPUs for very high resolution SR, while providing high quality of image enhancement. The core idea is to down-scale the input image to a lower resolution without losing information using a proposed pooling mechanism, hence the enhancement process is performed on a smaller scale, and the processing time is reduced. The enhanced data is then up-scaled back to the original resolution using sub-pixel interpolation.

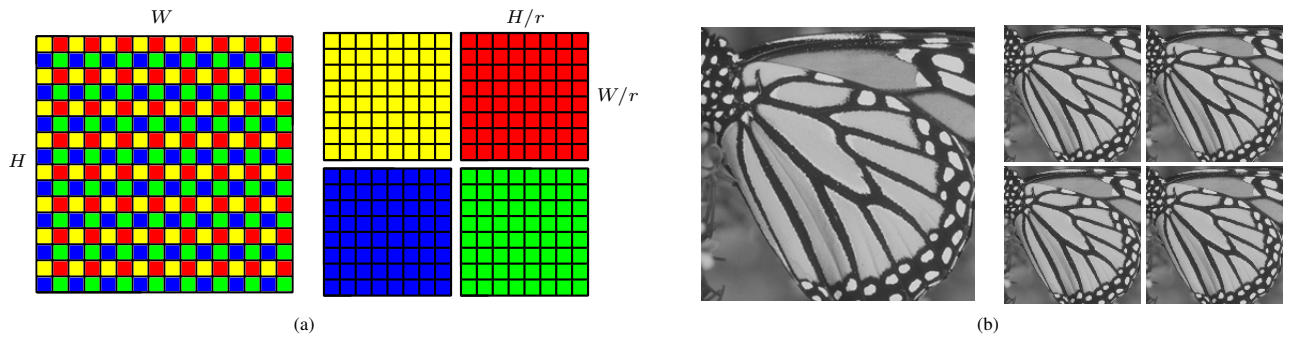


Fig. 2. The super-pixel pooling process: (a) Input and output of the layer for $r = 2$, and (b) a sample super-pixel pooling on a grayscale image.

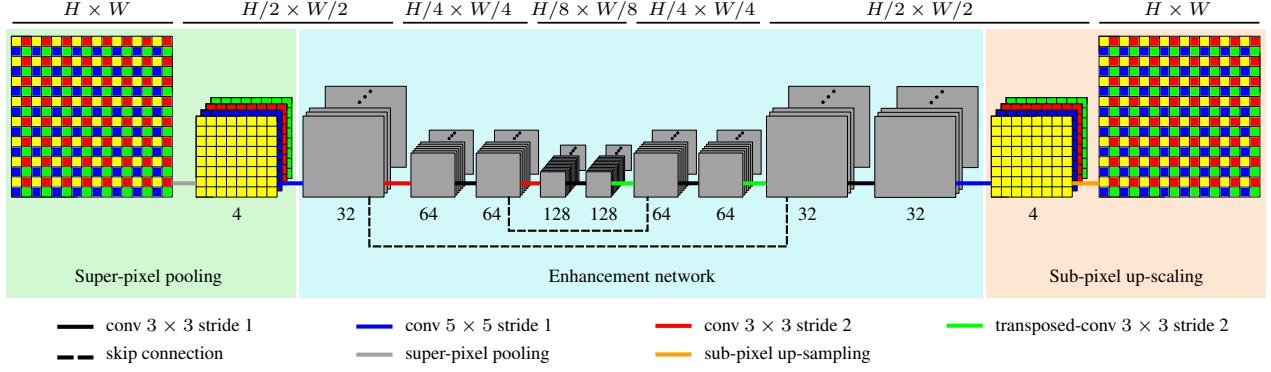


Fig. 3. The architecture of the proposed CNN for digital zoom enhancement using SR.

3.1. Super-pixel pooling layer

As mentioned earlier, the reason behind the high complexity of the image SR for high resolution scaling is simply the fact that all the convolutional operations need to be performed on very large matrices. Our proposed solution for complexity reduction is an initial layer for the CNN architecture called super-pixel pooling layer.

Normal pooling layers in CNNs result in the loss of data, however we propose a pooling layer that down-scales a single-channel image to a multi-channel image with lower resolution. This pooling mechanism is performed by rearranging a $H \times W$ matrix \mathbf{I} into a $\frac{H}{r} \times \frac{W}{r} \times r^2$ tensor \mathbf{O} . This operation can be considered as a reverse sub-pixel up-scaling defined in [1], and can be described mathematically as the following:

$$\mathbf{O}(x, y, c) = \mathbf{I}(r \cdot x - \text{mod}(r^2 - c, r), r \cdot y - \left\lfloor \frac{r^2 - c}{r} \right\rfloor) \quad (1)$$

where x , y , and c represent the coordinates of a pixel in \mathbf{O} . Although the above description aims at super-pixel pooling of single-channel (grayscale) images, the concept can be easily extended to multi-channel images. Fig. 2 demonstrates the super-pixel pooling operation, along with a visual example.

Application of super-pixel pooling layer reduces the spatial size of the input data without losing any information, and reduces the computation cost of convolution operation by a factor of r^2 . Moreover, the complexity of this pooling process is significantly lower than employing a convolution layer with a stride of r , which also results in a down-scaling.

3.2. Proposed network architecture

As mentioned earlier for enhancing the digital zoom functionality we require an SR framework which improves the quality of an image already up-sampled to the target resolution. We deployed an hourglass-shaped CNN with encoder-decoder structure inspired by the REDNet10 network presented in [9]. This network forms the basis of the enhancement framework.

We integrated the super-pixel pooling layer as an initial layer to the enhancement CNN to reduce the complexity. In order to reconstruct the high resolution image at the end of the process and achieve the target resolution, a sub-pixel up-scaling operation, as defined in [1], was applied. We chose $r = 2$ for the super-pixel pooling, and consequently the sub-pixel up-scaling layers. It is worth noting that selection of r is irrespective of the targeted SR scaling factor, and can be chosen completely independent of the enhancement network. Moreover, the presented architecture can be easily tuned to different scaling factors and different levels of enhancement. Additionally, the model can be trained simultaneously for several scaling factors at the same time, unlike other low complexity deep learning-based SR models, such as [1] and [4], that require separate models for different scaling factors. The end-to-end model can be formulated as the following:

$$\mathbf{X}^* = f(\mathbf{Y}, \boldsymbol{\theta}, r) \quad (2)$$

where \mathbf{Y} represents the low quality input image, up-scaled to the target resolution using an interpolation operation, \mathbf{X}^* represents the high quality version of the input image created by the network output, $\boldsymbol{\theta}$ represents all the CNN parameters including the kernels and biases within each layer, and r represents the scaling parameters for the super-pixel pooling and sub-pixel up-scaling layers.

Fig. 3 depicts the architecture of the proposed model for image enhancement. As illustrated, the network takes advantage of coupled convolutional and transposed convolutional layers, along with skip connections, that can lead to swift and accurate training of the CNN. Moreover, application of convolution layers with strides of higher than one leads to further complexity reduction along the enhancement work-flow. All the convolution and transposed convolution layers are accompanied by ReLU layers, and the kernel sizes are specified in Fig. 3. It is important to note that the presented SR approach is basically a complexity-reduced version of the REDNet10 model, which is essentially the core enhancement part of our model. In principle, REDNet10 is our proposed model without the super-pixel pooling and sub-pixel up-scaling steps.

3.3. Training

Training the proposed enhancement architecture is performed by solving an optimization problem to minimize the error between the labels and the CNN output. The training labels are a set of high resolution image samples \mathbf{X} , and the network outputs are the high quality high resolution image set generated by the model. The mean squared error, defined as the following, was employed as the cost function for the training process.

$$J_{MSE}(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}, r) = \frac{1}{N} \sum_{i=0}^N (\mathbf{X}_i - f(\mathbf{Y}_i, \boldsymbol{\theta}, r))^2 \quad (3)$$

It is worth noting that the training process can be done only for the enhancement section of the model, by performing super-pixel pooling on all the training samples as a pre-processing step prior to the training, so the training set would be compatible with the input and output structure of the enhancement component. This way a faster training process can be expected.

The training input samples are created by down-sampling the high resolution samples, and up-sampling them back to the original resolution by bi-cubic interpolation. The scaling factor can be fixed to focus on training a particular scaling, or alternatively can include several values to cover a wide range of sampling.

4. EXPERIMENTS

We focused the experiments on the SR with a scaling factor of 4, which is a challenging factor in image up-sampling. Moreover, since most of the mobile phone cameras provide a maximum $\times 4$ digital zoom, the experiments are compatible with the existing camera technologies.

We used the DIV2K data set [10], which comprises 800 high quality images as our training set. The images were par-

tioned into 96×96 samples with a stride of 80, which led to about 325,000 training sample pairs. The r parameter was also selected to be 2 for the experiments. We used Adam optimizer [11] for training the model with a learning rate of 0.0001 and a training batch size of 128. The proposed model, along with the existing state-of-the-art approaches, were implemented using TensorFlow¹ library.

4.1. Comparison with state of the art

To compare the presented model with the existing solutions, we used Set 5 and Set 14 data sets, which are widely used in SR task, as well as the Ultra-High-Definition (UHD) eye tracking data set [12], comprising 41 high quality images with 3840×2160 resolution. We compared the approach with well-known deep learning-based SR models including SRCNN [2], ESPCN [1], FSRCNN [4], and of course REDNet10 [9], which our work is based on. All the models were implemented and trained according to the provided information in the literature. All the trainings were performed on Tesla K80 NVIDIA GPUs, and all the tests were performed on a machine with a generic Intel Core i7-6700 CPU with a 3.40GHz clock and 16GB RAM.

Table 1 summarizes the quality performance of the presented model, in comparison with state-of-the-art methods. The quality metric is the Peak Signal-to-Noise Ratio (PSNR), which is widely used in SR evaluation. According to the results, the proposed approach can perform as good as the REDNet10, which is the baseline approach for this work, promising a high quality enhancement for images.

As the main focus of this work is devising a high quality SR with low complexity, it is important to compare the processing time of different approaches, too. Summarized in Table 2, the processing time for $\times 4$ up-sampling of an image to the UHD resolution using different methods shows that

¹<https://www.tensorflow.org/>

Table 1. PSNR analysis of the proposed method and state-of-the-art approaches for the scaling factor of 4.

	Bi-cubic	SRCNN	ESPCN	FSRCNN	REDNet10	Ours
Set 5	28.44	30.09	30.41	30.58	31.38	31.42
Set 14	26.00	27.18	27.37	27.52	27.98	27.95
UHD set	30.97	31.84	32.06	32.12	32.46	32.44
Average PSNR	28.47	29.70	29.95	30.07	30.61	30.60

Table 2. Complexity analysis (processing time in seconds) of the proposed method and state-of-the-art approaches for up-sampling of factor 4 to UHD resolution.

	SRCNN	ESPCN	FSRCNN	REDNet10	Ours
Processing time	121.01	0.76	1.10	151.93	5.62

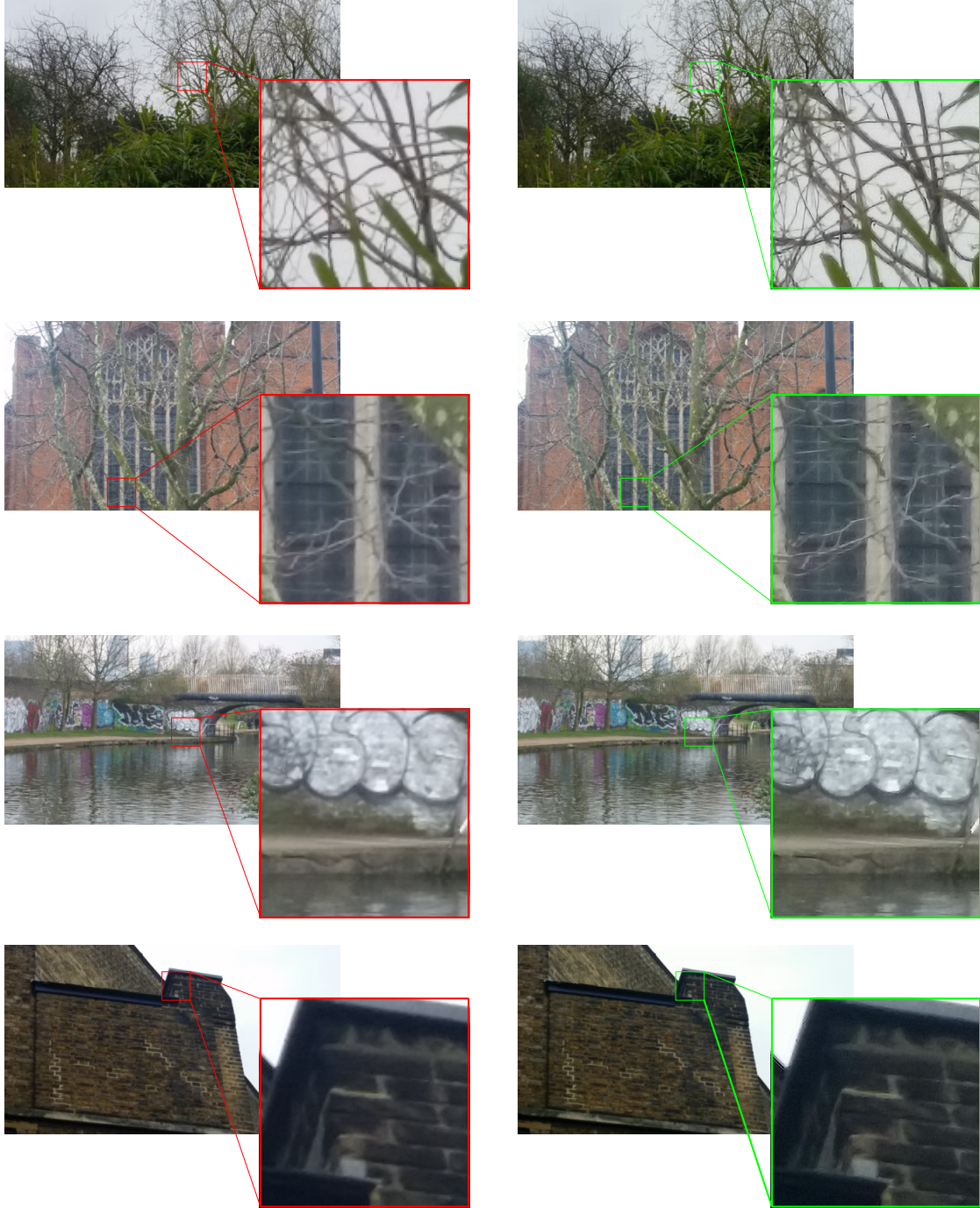


Fig. 4. Quality enhancement of digital zoom: The original captured images using Samsung Galaxy S5 with $\times 4$ digital zoom (left), and the enhanced images (right).

our approach reduces the complexity of REDNet10 by 96%, while providing the same image quality. This promises a very efficient performance on CPU devices, as well as easy integration in devices with moderate processing power, such as smart phones and tablets, as well as cloud systems. When compared with ESPCN and FSRCNN, although our approach

is still slower, the image quality is significantly better using the proposed method. Moreover, our method has a consistent structure and complexity regardless of the scaling factor, whereas in ESPCN and FSRCNN, the complexity of the structure is dependent on the scaling factor, which in both cases scaling up to a target resolution becomes slower, when

the scaling factor is lower. It is also worth noting that selecting a higher r parameter in our approach will result in further complexity reduction of the SR process.

4.2. Subjective evaluation

We also examined the subjective quality of the enhanced images when applied on real photos taken by mobile phones in maximum zoom-in mode. We used a Samsung Galaxy S5 phone, that has a 16-mega-pixel camera, and allows a $\times 4$ digital zoom. We tested the camera with maximum zoom, which results in photos that are interpolated to the 5312×2988 resolution. Application of the proposed enhancement approach resulted in clear visual improvements of the photos, some of which are presented in Fig. 4. In these examples only the luma signal is enhanced by the proposed SR-based enhancement approach.

5. CONCLUSIONS

We presented a deep learning-based SR approach for enhancement of the digital zoom in mobile phone cameras that takes advantage of a super-pixel pooling layer. Application of this pooling layer leads to major speed-ups in processing and enhancement of high resolution images, while providing a high visual quality comparable with the baseline approaches. The presented CNN architecture can be trained for several scaling factors simultaneously, and the complexity is independent of the scaling factor unlike existing low complexity SR models. It is also worth mentioning that the presented SR method is also applicable for generic spatial up-sampling of still images and videos regardless of the application.

6. ACKNOWLEDGMENTS

The work described in this paper has been conducted within the project COGNITUS. This project has received funding from the European Unions Horizon 2020 research and innovation program under grant agreement No 687605. This research utilized Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT.

7. REFERENCES

- [1] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1874–1883.
- [2] 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 (ECCV)*, 2014, pp. 184–199.
- [3] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
- [4] Chao Dong, Chen Change Loy, and Xiaoou Tang, "Accelerating the Super-Resolution Convolutional Neural Network," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 184–199.
- [5] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.
- [6] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Deeply-Recursive Convolutional Network for Image Super-Resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.
- [7] Zhen Cui, Hong Chang, Shiguang Shan, Bineng Zhong, and Xilin Chen, "Deep Network Cascade for Image Super-resolution," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 49–64.
- [8] Zhangyang Wang, Yingzhen Yang, Zhaowen Wang, Shiyu Chang, Wei Han, Jianchao Yang, and Thomas S. Huang, "Self-Tuned Deep Super Resolution," in *CVPR Workshops*. 2015, pp. 1–8, IEEE Computer Society.
- [9] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang, "Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 2802–2810.
- [10] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.
- [11] Diederik Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations*, Dec. 2014.
- [12] Hiromi Nemoto, Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi, "Ultra-eye: Uhd and hd images eye tracking dataset," in *6th International Workshop on Quality of Multimedia Experience (QoMEX)*, Sep. 2014.