A Dual-Network Based Super-Resolution for Compressed High Definition Video

Longtao Feng[1,3(✉)], Xinfeng Zhang[2], Xiang Zhang[3], Shanshe Wang[3], Ronggang Wang[1], and Siwei Ma[3]

[1] Peking University Shenzhen Graduate School, Shenzhen, China
ltfeng@pku.edu.cn, rgwang@pkusz.edu.cn
[2] University of Southern California, Los Angeles, CA, USA
xinfengz@usc.edu
[3] Institute of Digital Media, Peking University, Beijing, China
{x_zhang,sswang,swma}@pku.edu.cn

Abstract. Convolutional neural network (CNN) based super-resolution (SR) has achieved superior performance compared with traditional methods for uncompressed images/videos, but its performance degenerates dramatically for compressed content especially at low bit-rate scenario due to the mixture distortions during sampling and compressing. This is critical because images/videos are always compressed with degraded quality in practical scenarios. In this paper, we propose a novel dual-network structure to improve the CNN-based SR performance for compressed high definition video especially at low bit-rate. To alleviate the impact of compression, an enhancement network is proposed to remove the compression artifacts which is located ahead of the SR network. The two networks, enhancement network and SR network, are optimized stepwise for different tasks of compression artifact reduction and SR respectively. Moreover, an improved geometric self-ensemble strategy is proposed to further improve the SR performance. Extensive experimental results demonstrate that the dual-network scheme can significantly improve the quality of super-resolved images/videos compared with those reconstructed from single SR network for compressed content. It achieves around 31.5% bit-rate saving for 4 K video compression compared with HEVC when applying the proposed method in a SR-based video coding framework, which proves the potential of our method in practical scenarios, e.g., video coding and SR.

Keywords: Super-resolution · Enhancement network
Compression artifact reduction · Video coding · HEVC
Convolutional neural network

1 Introduction

Due to fast development of the image/video capture and display technologies, the ultra high-definition (e.g., 4 K) content has becoming more and more popular. Increasing the image resolution to 4 K or higher will dramatically improve

R. Hong et al. (Eds.): PCM 2018, LNCS 11164, pp. 600–610, 2018.
https://doi.org/10.1007/978-3-030-00776-8_55

the user experience by leading to a more immersive view environment. However, the data size increases significantly at the same time, which makes new compression strategy for 4 K content important and indispensable. An efficient way for 4 K content compression is based super-resolution (SR), where the original video is downsampled before compression and the decoded video is upsampled to the original resolution using SR technologies. Besides compression, SR technologies are also demanded to display low resolution video onto high definition devices. However, traditional image interpolation methods cannot get visually satisfied results especially for compressed low resolution video and may incur blurring artifacts. Therefore, the learning based SR approaches have been widely investigated recently.

A+ [16] is one representative SR method in recent years using regression to learn the correlation between low-resolution (LR) and high-resolution (HR) patches, which combines the best qualities of Anchored Neighborhood Regression (ANR) [15] and Simple Functions (SF) [18] adaptively. Recently, the convolution neural network (CNN) based SR methods [1,2,4,6–8,11,17] have achieved significant improvement compared with the traditional methods. In [1], Dong et al. proposed the shallow convolution network, SRCNN, which achieves significant quality improvement against its previous methods. To optimize the SRCNN, Dong et al. further proposed a compact hourglass-shape CNN structure, FSRCNN [2], which is faster than SRCNN and achieves better performance. To further improve the SR performance, Kim et al. proposed a very deep convolution network by cascading many small filters, named VDSR [4]. In [8], Lim et al. developed an enhanced deep super-resolution network (EDSR) by removing unnecessary modules of Ledig et al.'s conventional residual networks [6] and won the NTIRE2017 Super-Resolution Challenge [14].

However, the above SR methods are based on uncompressed images/videos without considering compression influence. In practice, the available images and videos are all compressed versions, and the compression artifacts, e.g., blocking and ringing artifacts which have been studied for reduction in in-loop filters [3,20], can dramatically degenerate the performance of SR methods especially at low bit-rate scenario. This is essentially because of the mixture of two different degenerations, i.e., sampling and compressing. In addition, most of the SR approaches are investigated and verified on LR images and videos, e.g., 256×256, and there is little work for high-definition (HD) videos e.g., 1080P, which are assumed enough for display in the past years. Along with the wide deployment of ultra high-definition display devices, high efficiency SR algorithms for HD video is also urgently demanded for both display and compression applications.

In this paper, we focus on the compressed video SR from HD to 4 K resolution, and propose an end-to-end CNN method to optimize the quality of the super-resolved video by removing the compression artifacts, enhancing video resolution respectively and utilizing the improved geometric self-ensemble. Specifically, We divide the compressed video SR problem into two subtasks, i.e., video enhancement and video SR, which are solved by neural network methods. An enhancement network without pooling layers are proposed and located ahead

(a) Origin (b) Bicubic (c) VDSR

Fig. 1. Subjective results when applying SR methods to a compressed video frame.

of the SR network to reduce the compression artifacts firstly, and the SR network is then applied to obtain reconstructed 4 K video. Moreover, we also apply the proposed method to a SR based video compression framework, where the 4 K video sequences are first downsampled into 1080P and compressed at the encoder side, then the proposed SR method is applied to the decoded 1080P videos. Extensive experimental results show that the proposed SR method in SR based compression can achieve about 31.5% bit-rate saving compared with the latest video coding standard, High Efficiency Video Coding (HEVC) [13].

The rest of the paper is organized as follows: in Sect. 2, we introduce the proposed video SR method and the SR based video coding framework using the proposed method. The experimental results are shown in Sect. 3, and Sect. 4 concludes the paper.

2 Proposed Method

2.1 Motivation

At low bit-rate scenario, video compression introduces obvious compression artifacts, e.g., twisted lines, blurred edges and fuzzy textures, which are mainly caused by coarse quantization. In video coding, as the quantization parameter (QP) increases, more quantization noise will be introduced reducing the quality of the reconstructed video. Traditional SR methods cannot handle these compression distortions well and their performance degenerates seriously due to the severe compression distortions. Considering the compression process, the degeneration process of a compressed low resolution image y_d can be modeled as follows,

$$y_d = y \otimes s \otimes c, \tag{1}$$

where y represents the origin image, s and c denote sampling and compressing degeneration, respectively. The mixture of distortions cannot be resolved easily by single neural network due to the essentially different degeneration kernels. To show our motivation, we apply SR methods to compressed images directly. From Table 1, we can see that the performance of the bicubic interpolation and VDSR is poor on the HEVC compressed images, especially at low bit-rate scenario

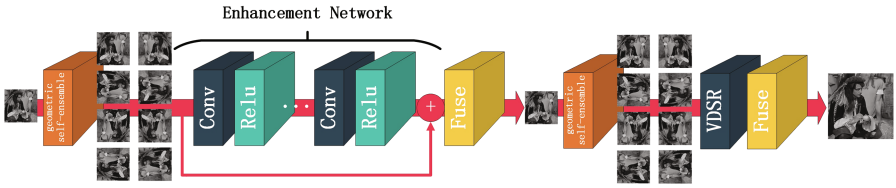


Fig. 2. The pipeline of the proposed SR method, where a dual-network structure is introduced. An enhancement network is applied before VDSR for compensating the compression distortions.

corresponding to high QPs. In Fig. 1, we further show a subjective result for a video frame compressed by HEVC at QP = 44. The corresponding network is trained based on the compressed images/videos. However, we can find that the compression distortions cannot be eliminated and are even enlarged after SR, e.g. the blocking artifacts around the face.

This motivates us to improve the SR performance by introducing an enhancement network before SR to reduce the compression artifacts and to solve them separately. Such dual-network structure is superior to cascading into one single network. Compared with tuning a highly deep network, a stepwise training is more feasible and efficient especially when the training set is limited.

2.2 The Proposed Video SR Method

According to the above analysis, we proposed a novel video SR method by adding an enhancement network before SR to reduce compression artifacts as shown in Fig. 2. The method can be modeled as follows,

$$\hat{y}_r = f_s(f_e(y_d)), \tag{2}$$

where $\hat{y}_r$ represents reconstructed high resolution image, f_e denotes enhancement operation which resolves compression degeneration and f_s denotes SR operation. Both enhancement and SR can be regarded as regression problems which aim to restore the high quality video from its distorted or LR version, but their degradation models are completely different. To solve the compression degeneration problem, we design a neural network with 20 convolutional layers taking the rectified linear unit (RELU) [9] as the activation function. It is inspired by the work of Kim et al. [4] which shows that the CNNs are efficient in dealing with regression problems. For SR problem, we adopt the network structure of VDSR [4] due to its high efficiency and good performance.

Since the two networks aim for different degradation models, we train them separately. The enhancement network is first trained. Given a training dataset $\{x_c^i, y_{uc}^i\}_{i=1}^N$, x_c^i represents a compressed image/video and y_{uc}^i is the corresponding uncompressed version. Our goal is to learn a model f_e which restores a image/video from its compressed version: $\hat{y_e} = f_e(x_c)$, where $\hat{y_e}$ is the restored

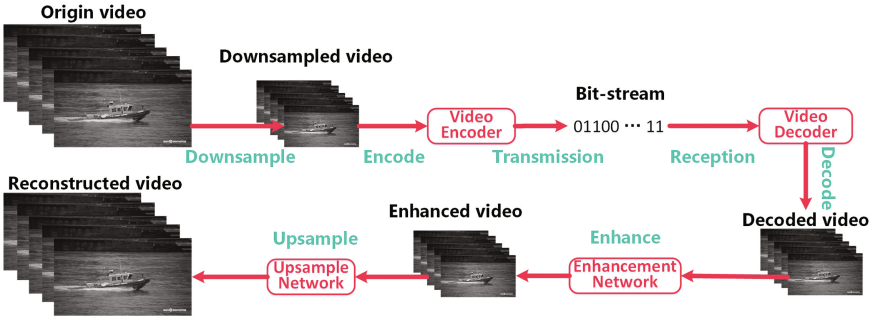


Fig. 3. The SR based video compression framework.

version of a compressed image/video. The mean squared error (MSE) is utilized as the loss function of the enhancement network:

$$L_e = \frac{1}{2}||y_{uc} - \hat{y_e}||^2. \tag{3}$$

Then, we use the existing VDSR method [4] as the SR network, of which the input is the output of the enhancement network. The data for training can be denoted as $\{x_e^i, y_h^i\}_{i=1}^N$, where x_e^i represents a LR compressed image/video which is enhanced by the enhancement network and y_h^i represents the corresponding uncompressed HR image/video. The proposed method in Fig. 2 is a flexible framework that leaves the choice for a specific network architecture open. Our choice of the network architectures provides a solution for the tradeoff between performance and complexity. More recent methods for each subtask, especially more complex SR methods [6,8] can be easily incorporated and will lead to even better results.

Moreover, to further improve the performance, the geometric self-ensemble strategy [8] is adopted and modified for our problem. Specifically, each input video I^{input} is flipped vertically and rotated to generate seven augmented inputs $I_{n,i}^{input} = T_i(I_{n,1}^{input})$, where T_i represents the i^{th} geometric transformations including identity, i.e., $i = 1, \cdots, 8$. With those augmented input videos, we can generate the corresponding processed videos $\{I_{n,1}^{output}, ..., I_{n,8}^{output}\}$. Finally, we can generate the output video by inversely transforming the 8 processed video frames to their original structures and fusing them by an introduced weighting factor as follows,

$$I_{n,i}^{output} = \alpha \tilde{I}_{n,1}^{output} + \frac{(1-\alpha)}{7} \sum_{i=2}^{8} \tilde{I}_{n,i}^{output} \tag{4}$$

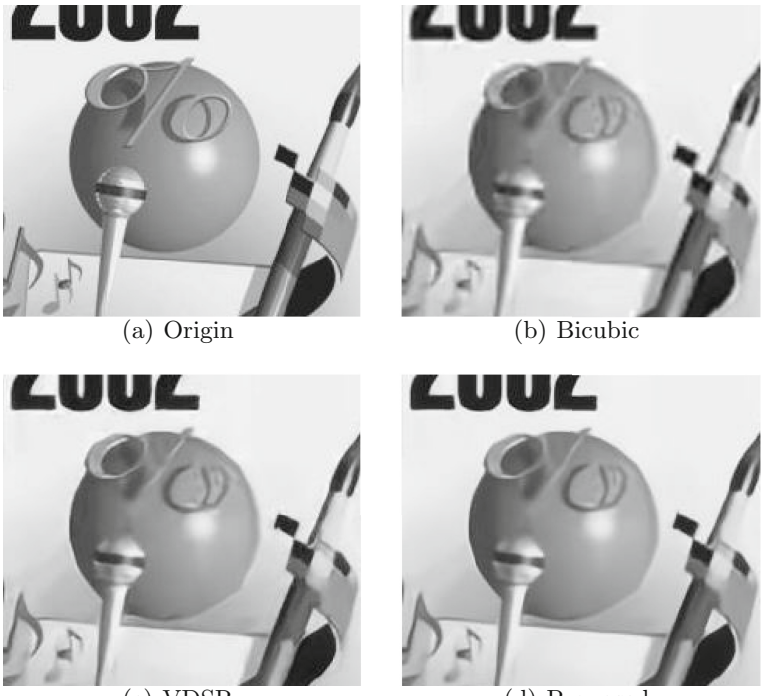
where α is set to 0.3, which is based on the assumption that the output from identity video is closer to the original geometric structure than those from transformed inputs.

Table 1. PSNR of different SRs as a function of QP on a image testset.

Set14	Bicubic	VDSR	Proposed
Uncompressed	29.63	32.45	–
QP34	27.68	28.60	28.79
QP38	26.56	27.19	27.37
QP44	24.48	24.90	25.02

2.3 The Application in SR Based Video Compression

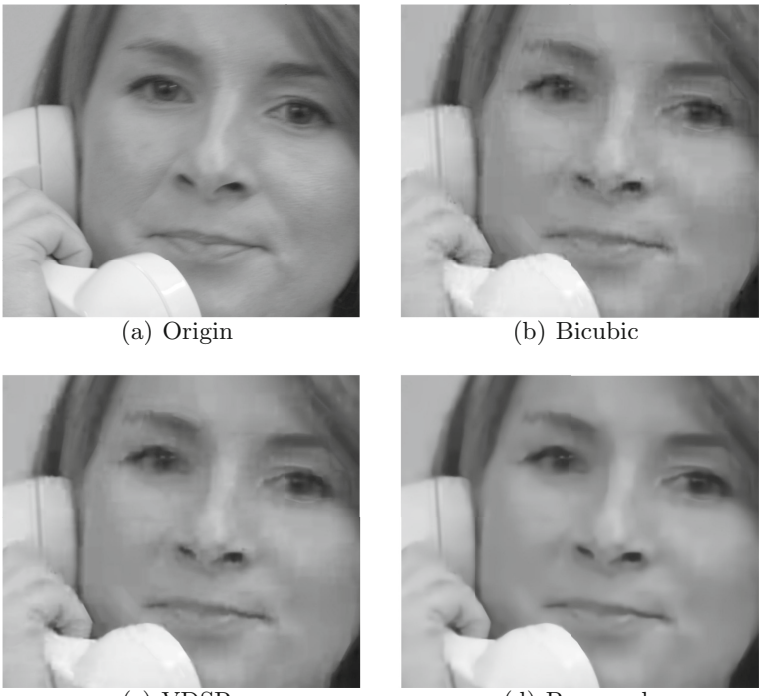
Limited by the bandwidth, the 4 K video is usually compressed at low bit-rate, which has obvious compression artifacts. One efficient solution is the SR based video compression strategy, where the videos are first downsampled before compression and get upsampled at the decoder. Therefore, the proposed method can be naturally applied to this application because the dual-network can help in reducing the compression artifacts. The introduced compression framework is shown in Fig. 3.



(a) Origin (b) Bicubic

(c) VDSR (d) Proposed

Fig. 4. Subjective comparison with different SR methods for images.



(a) Origin (b) Bicubic

(c) VDSR (d) Proposed

Fig. 5. Subjective comparison with different SR methods for videos.

3 Experimental Results

3.1 Datasets and Parameters

For training, we first use 291 images as in [10] to train models for initialization. Since ultra high-definition video has different characteristics compared with low-resolution one, we fine-tuned the dual-network using 4 K video sequences further, where the SJTU 4K video dataset in [12] and some other 4 K video sequences collected by ourselves are utilized. To avoid repeatability, only one frame for each video is extracted as training data and its flipped version is also added as augmentation training data. At the test stage, we use two dataset, including the

Table 2. PSNR of different SRs as a function of QP on video test set.

AWS	QP44				QP38				QP34				QP30			
	Bicu-bic	VDSR	Pro-I	Pro-II	Bicu-bic	VDSR	Pro-I	Pro-II	Bicu-bic	VDSR	Pro-I	Pro-II	Bicu-bic	VDSR	Pro-I	Pro-II
Cactus	35.33	35.84	36.11	36.37	39.13	39.85	39.94	40.10	41.07	42.20	41.65	42.53	42.46	43.92	43.10	44.13
Coastguard	30.29	30.45	30.62	30.63	33.02	33.27	33.39	33.43	35.02	35.34	35.41	35.54	37.08	37.49	37.48	37.65
Foreman	35.12	35.48	35.74	35.80	38.24	38.75	38.83	38.94	40.04	40.64	40.57	40.89	41.59	42.30	42.08	42.46
News	35.24	35.87	35.88	36.28	38.52	39.68	39.20	39.88	40.46	42.09	41.06	42.49	42.04	44.15	42.57	44.53
Suzie	35.20	35.62	35.98	36.11	38.02	38.59	38.75	38.72	39.73	40.37	40.32	40.55	41.30	41.90	41.76	42.04
Average	34.23	34.65	34.86	35.04	37.39	38.03	38.02	38.21	39.26	40.13	39.80	40.40	40.89	41.95	41.40	42.16

image test set "set14" in [19] which are widely utilized as benchmark for SR [1,15,16] and the video test set with five 4 K video sequences released by AWS elemental[1]. It is worth noting that for videos the SR methods are applied to each video frame. These images and videos are downscaled by bicubic interpolation and compressed by HEVC reference software, HM 16.17, under low-delay-P configuration in both training and testing stages.

We train our model using ADAM optimizer [5]. The training is regularized by weight decay (ℓ_2 norm penalty multiplied by 0.0001). The minibatch size is set as 64. The learning rate and training epoch vary with QPs and different types of network. More details can be found in our Github repository[2].

3.2 Experimental Results

For evaluation, we compare our model with the baseline method: the VDSR SR method trained by compressed images, the equivalent of our propose method without enhancing. In this work, we only process the luminance component for all methods, because human vision is more sensitive to details in intensity than that in chroma components.

For image results, Table 1 denotes the PSNR of the upsampled HR images from compressed images using different methods. It shows that performance degenerates with quantization when using baseline due to the mixture of two degeneration and our proposed method can compensate the losses caused by compression. Figure 4 shows one example of subjective results for different methods. We can find that the image quality is improved and some compression distortions, e.g., the ringing and blocking artifacts, are suppressed by our method.

For video results, Table 2 shows the PSNR results of upsampled HR video sequences from compressed videos using different methods. To further evaluate how the proposed enhancement network perform on different SR schemes, we compare with two methods, the Pro-I and Pro-II, respectively. The Pro-I is essentially the enhancement plus bicubic interpolation, and Pro-II is essentially the enhancement plus VDSR. From the results, we can achieve several conclusions. First one can be observed that both Pro-I and Pro-II can increase the PSNR indicating the effectiveness of the enhancement network. Next, it is clear that the SR method itself will also impact on the ultimate results, where an advanced SR will bring more gains. Third, We can see that the proposed method achieves superior results compared with either Bicubic or VDSR only. More importantly, the proposed method can obtain more PSNR gains for lower bit-rate, where about 0.4 dB gain can be achieved when QP = 44. This further implies that the proposed enhancement network is indispensable in improving the SR performance for compressed videos. Figure 5 compares the subjective results of different methods. Similar observations can be achieved that the video quality is improved and the compression distortions can be reduced by our method, e.g. the blocking artifacts around the face.

[1] http://www.elementaltechnologies.com/resources/4K-testsequences.
[2] https://github.com/FLT19940317/supplementary-material-of-PCM2018-Paper.

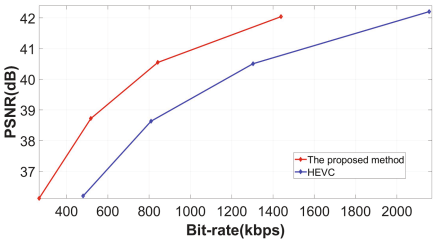


Fig. 6. RD curve comparison between the proposed and HEVC for sequence *Suzie*.

Table 3. The BD-rate of SR based compression using the proposed method compared with HEVC at low bit-rate.

AWS Sequence	BD-rate Y(%)
Cactus	−33.91
Coastguard	−24.31
Foreman	−33.86
News	−28.27
Suzie	−37.07
Average	−31.48

For the video compression application, Table 3 shows bit-rate saving compared with HEVC according to the QPs in Table 2 when applying the proposed SR method to the SR based video coding framework as shown in Fig. 3. The test sequences in Table 3 are all 4 K video sequences. Obviously the proposed SR method benefits the SR based video compression, which achieves over 30% bit-rate saving on average compared with HEVC.

Furthermore, in Fig. 6, we show the rate-distortion (RD) curves of the proposed method and HEVC according to the QPs in Table 2 for video sequence *Suzie*. We can see that the proposed SR based video compression outperforms HEVC significantly in a relative large bit-rate range. Therefore, the proposed dual-network based SR is efficient in improving the SR performance for compressed content.

4 Conclusion

In this paper, we present a dual-network based super-resolution (SR) method for compressed content, where an enhancement network is introduced before a SR network. The proposed method resolves two different degenerations stepwise and achieves better performance on SR task compared with existing methods, indicating the efficiency of the proposed scheme. In addition, the proposed method also shows its advantage in SR based video coding application, and significant bit-rate saving is obtained compared with HEVC. In future work, we will investigate the inter-frame correlations in both the enhancement and SR networks.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (61571017), National Postdoctoral Program for Innovative Talents (BX201600006)Top-Notch Young Talents Program of China, High-performance Computing Platform of Peking University, which are gratefully acknowledged.

References

1. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
2. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 391–407. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_25
3. Jia, C., Wang, S., Zhang, X., Wang, S., Ma, S.: Spatial-temporal residue network based in-loop filter for video coding. In: 2017 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4. IEEE (2017)
4. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1646–1654 (2016)
5. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
6. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint (2016)
7. Liang, Y., Timofte, R., Wang, J., Gong, Y., Zheng, N.: Single image super resolution-when model adaptation matters. arXiv preprint arXiv:1703.10889 (2017)
8. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, vol. 1, p. 3 (2017)
9. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the ICML, vol. 30, p. 3 (2013)
10. Schulter, S., Leistner, C., Bischof, H.: Fast and accurate image upscaling with super-resolution forests. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3791–3799 (2015)
11. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883 (2016)
12. Song, L., Tang, X., Zhang, W., Yang, X., Xia, P.: The SJTU 4K video sequence dataset. In: 2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 34–35. IEEE (2013)
13. Sullivan, G.J., Ohm, J., Han, W.J., Wiegand, T.: Overview of the high efficiency video coding (hevc) standard. IEEE Trans. Circuits Syst. Video Technol. **22**(12), 1649–1668 (2012)
14. Timofte, R., et al.: Ntire 2017 challenge on single image super-resolution: methods and results. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1110–1121. IEEE (2017)
15. Timofte, R., De, V., Van Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 1920–1927. IEEE (2013)
16. Timofte, R., De Smet, V., Van Gool, L.: A+: adjusted anchored neighborhood regression for fast super-resolution. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9006, pp. 111–126. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16817-3_8

17. Wang, Y., Wang, L., Wang, H., Li, P.: End-to-end image super-resolution via deep and shallow convolutional networks. arXiv preprint arXiv:1607.07680 (2016)
18. Yang, C.Y., Yang, M.H.: Fast direct super-resolution by simple functions. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 561–568. IEEE (2013)
19. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Boissonnat, J.-D., et al. (eds.) Curves and Surfaces 2010. LNCS, vol. 6920, pp. 711–730. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-27413-8_47
20. Zhang, X., Wang, S., Zhang, Y., Lin, W., Ma, S., Gao, W.: High-efficiency image coding via near-optimal filtering. IEEE Signal Process. Lett. **24**(9), 1403–1407 (2017)