CNN Based Multi-view Image Quality Enhancement

Gyu-Lee Jeon, Hee-Jae Kim, Eun Yeo, Je-Won Kang
Department of Electronic and Electrical Engineering and Graduate Program in Smart Factory,
gyulee0412@gmail.com, heejaekim@ewhain.net, eun.yeo@ewhain.net, jewonk@ewha.ac.kr

***Abstract*— We propose a CNN-based multi-view image quality enhancement (MVIQE) to improve the quality of a target image using adjacent multi-view images. Differing from the conventional single frame quality enhancement, our approach aims to improve the quality by transferring a higher quality of other input views in multi-view images as references. Our network contains an optical flow estimation module, warping layers, and image synthesis module to enhance the quality of a target image with quantization noise. Experimental results show that our method outperforms previous studies on image quality enhancement in terms of peak signal-to-noise ratio performance.**

Keywords—Convolutional Neural Networks(CNN), Image quality enhancement, multi-view images

I. INTRODUCTION

Recently, multi-view video contents are widely used with three-dimensional (3D) immersive video services such as metaverse and virtual reality. However, the image quality of the compressed video drops rapidly with limited bandwidths [17]. A video compression rate tends to increase in 3D video transmission due to a number of views, and accordingly, image quality needs to be further enhanced.

Convolutional neural network (CNN)-based techniques are widely used to improve degraded image quality by quantization noise (Q-noise) during video compression [1, 18]. In the process of video compression, the quantization brings in distortion because the quantization determines how much high-frequency signal will be lost during compression. The compression rate is controlled by a quantization parameter (QP). The higher the QP value, the more high-frequency signals are lost. Then, the image has poor quality.

Increasing the quality of the compressed image makes it possible to visually recognize an object more accurately. Although there were several studies to alleviate the degradation of an image quality [2, 3], it is difficult to directly apply those methods to multi-view images [19].

In this paper, we propose a method to increase the quality of a low-quality compressed target image by using its high-quality multi-view images as reference frames. The main contributions of this paper are: (1) we propose a novel CNN-based multi-view image enhancement, which can reduce compression artifacts of a compressed image using adjacent views among input multi-view images. (2) We develop a new fusion scheme utilizing multi-view image characteristics. (3) Experimental results demonstrate that the proposed method outperforms state-of-the-arts studies in image enhancement.

II. RELATED WORK

A. Image quality enhancement

Image quality enhancement research is being conducted to enhance the image quality that has degraded quality due to compression. In the case of JPEG compression, blocky or ringing artifacts occur when compressing an image. To solve these side effects, deep learning-based compressed image quality enhancement research is in progress. For example, there are AR-CNN [2] and Dn-CNN [3]. However, all of the above methods are single image enhancement methods, which are not efficiently applicable to multi-view images. Because the existing single image enhancement methods do not utilize the characteristics of the other images in different views, the degree of quality improvement for multi-view images could be further improved.

B. Super-resolution

In the compressed image enhancement, although only few studies have been conducted to enhance the quality of a target image using multi-view images as reference images, there were several pioneering studies on reference-based super-resolution (Ref-SR) as a pixel-based restoration task [4, 5]. Ref-SR is an image processing task to increase the resolution of a low-resolution image to a high-resolution by using the high-resolution image as a reference image. Since a reference image has high-resolution information, if it is used to super-resolve the low-resolution target frame, as a result, an image with clean quality can be obtained. Accordingly, the Ref-SR studies provided superior performance to single-image-based SR [6, 7].

State-of-the-arts reference-based super-resolution studies so far include CrossNet [8] and TTSR [9]. CrossNet is for light field image super-resolution, which does not consider characteristics of multi-view images. Specifically, in light field images, the disparity is smaller than in that of multi-view images. Although it is easier to estimate flows of light-field images, it cannot be directly applied to multi-view images. TTSR uses the texture of a reference image for super-resolution. However, this method is not fitted for compressed images because the compressed image has already lost fine textures. It would be difficult to use the textures of reference images for improving the target compressed image. Therefore, for considering the

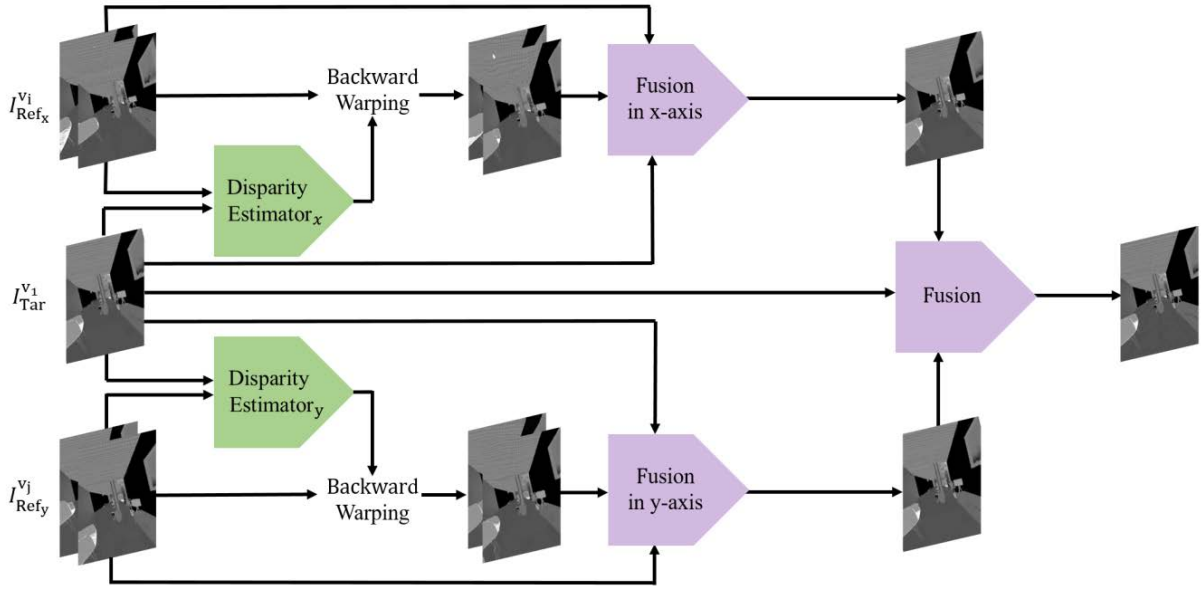

Figure 1. Network structure of our MVIQE, including optical flow estimation, backward warping layers, and image synthesis.

characteristics of compressed multi-view images, we propose a novel multi-view compressed image quality enhancement.

C. Image/video Synthesis using Warping

Our method is related to image/synthesis tasks that use additional images from other viewpoints or images. The synthesis methods include view synthesis [10] and video interpolation [11, 16]. However, our method is different from the existing synthesis methods. We devised a new fusion scheme to take advantage of the multi-view images' characteristics.

III. Proposed Approach

This model aims to enhance the quality of a compressed target image $I^{v_1}_{Tar}$ by using multi-view images $I^{v_i}_{Ref_x}$, $I^{v_j}_{Ref_y}$ as its references captured in views of v_i and v_j along with x -axis and y -axis. It is noted that the reference images are high-quality images compressed with a low QP, and the target image is a low-quality image compressed with a high QP. In other words, we use high-quality references for improving target quality. Our model architecture is pipelined to the optical flow estimation, backward warping layers, and image synthesis. The entire network architecture is plotted in Fig. 1.

A. Alignment module and Warping loss

Since the reference multi-view images are not aligned with the target image, we align references to the target using flow estimation and backward warping layers. We employ PWC-Net [12], which provides high accuracy to produce flows, to estimate the disparity between target and references. PWC-Net estimates a flow F_i between each reference view image $I^{v_k}_{Ref}$ and the target view image $I^{v_1}_{Tar}$ as follows:

$$F_i = flow\left\{I^{v_1}_{Tar}, I^{v_k}_{Ref}\right\}, \quad (1)$$

where $flow$ refers to a network function of PWC-Net. In experiments, we use four flows from four references, i.e., i=1..4 and k=2..5.

After that, with each flow F_i, in backward warping layers, all reference images $I^{v_k}_{Ref}$ are aligned to the same view point of a target image $I^{v_1}_{Tar}$ in the image domain as follows:

$$I_{warped} = Backward\ Warp\{F_i, I^{v_k}_{Ref}\} \quad (2)$$

where $Backward\ Warp$ refers to the backward warping layer in Fig. 1 to produce an output reference images I_{warped}.

However, warping between the target image and reference multi-view images is difficult due to a large disparity of the multi-view images. Therefore, for accurate warping, we propose to add a warping loss L_{warp}. In the case of multi-view images, since the disparity difference between each view is large, a warping loss is essential. After adding the warping loss function, this disparity between target and references can be estimated well. This warping loss L_{warp} refers to the difference between the ground-truth image I_{gt} and warped reference images I_{warped}. By adding this warping loss, the model can continue to estimate an optical flow well during training for alignment from reference images to the target image. Given the network prediction $I_{enhanced}$, and the ground-truth image I_{gt}, the loss function can be written as

$$L_{recon} = \sqrt{\| I_{enhanced} - I_{gt} \|^2 + \rho^2}, \quad (3)$$

and

$$L_{warp} = \sqrt{\| I_{warped} - I_{gt} \|^2 + \rho^2}, \quad (4)$$

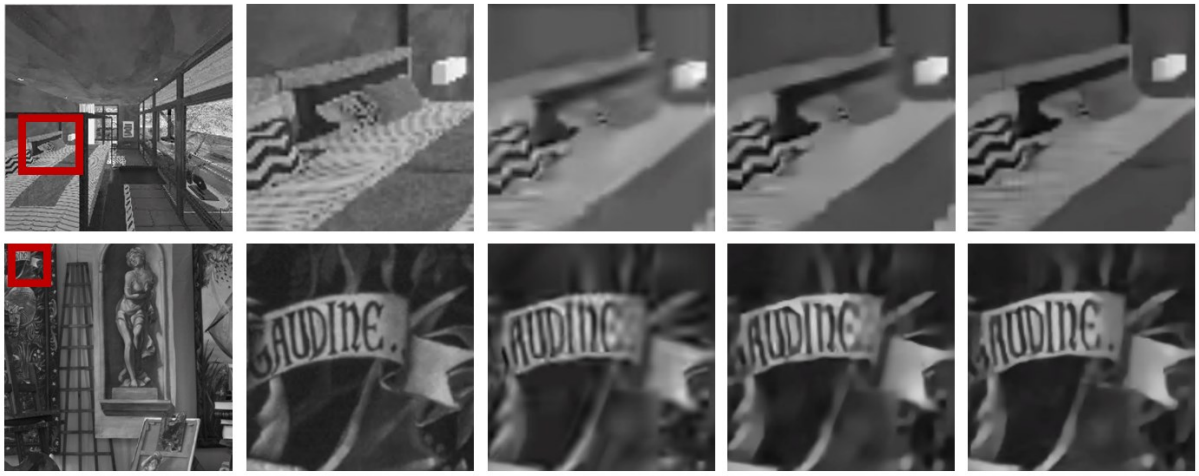

Figure 2. Qualitative evaluation comparisons

where $\rho = 0.001$ is the Charbonnier penalty function [13].

We build the total loss function as follows:

$$L_{total} = L_{recon} + L_{warp} \quad (5)$$

B. Fusion module

Multi-view images are obtained on the x-axis or y-axis around the target view. We propose a new fusion scheme using this characteristic. First, we attempt to warp all of the reference views to the target view. Then, for the x-axis, we try to fuse three kinds of features from the target view $I_{Tar}^{v_1}$, the reference views $I_{Ref_x}^{v_2}$ and $I_{Ref_x}^{v_3}$, and the warped reference views $I_{warped_x}^{v_2}$ and $I_{warped_x}^{v_3}$. We concatenate these images along the x-axis and then put the images into a reconstruction module [7] to obtain a new synthesis image. We call this new fusion image the x-axis fusion image I_{fusion_x} given as.

$$I_{fusion_x} = rec\{ I_{Ref_x}^{v_2} + I_{warped_x}^{v_2} + I_{Tar}^{v_1} + I_{warped_x}^{v_3} + I_{Ref_x}^{v_3} \} \quad (6)$$

where rec represents the reconstruction module. In eq(6), the input images have consistency along the x-axis as in interpolated video frames. Similarly, it creates the y-axis fusion image I_{fusion_y}, given as

$$I_{fusion_y} = rec\{ I_{Ref_y}^{v_4} + I_{warped_y}^{v_4} + I_{Tar}^{v_1} + I_{warped_y}^{v_5} + I_{Ref_y}^{v_5} \} \quad (7).$$

Finally, the target view image $I_{Tar}^{v_1}$, x-axis fusion image I_{fusion_x}, and y-axis fusion image I_{fusion_y} are synthesized to generate enhanced output $I_{enhanced}$.

$$I_{enhanced} = rec\{I_{Tar}^{v_1} + I_{fusion_x} + I_{fusion_y}\} \quad (8)$$

This fusion method is using the characteristics of the multi-view image. By using this method, the disadvantage of multi-view, which has a large disparity and a large difference between the target view and multi-views, can be reduced. That is, the multi-view features are easy to be used for improving the target image thanks to consistent alignment.

IV. Experiment

A. Datasets

The dataset used in the experiment is generated from synthetic ERP datasets [4, 5] and converted to perspective synthetic data with 6 scenes. It has 5 views based on one scene. The first view v_1 is selected as the target view, and the remaining 4 views (v_2-v_5) are selected as the reference multi-view. For one scene, images generated in chronological order as synthetic data are converted into YUV video format. The YUV video is encoded with QP = 32, 37(for references), QP = 37, 42(for target) using High Efficiency Video Coding (HEVC) [1] reference software under all intra configuration. The quality difference between target view and reference views are around 3dB. The dataset consists of 4,800 images in total, 2,880 images for training, 960 images for validation, and 960 images are used for testing. All image size is 256 x 256.

To test the generalization ability of our method, we also test on the images from several multi-view dataset used in Moving Picture Expert Group (MPEG) 3D video coding standardization [14]. The dataset is captured from natural contents with perspective view. This dataset is also encoded and then used to test the performance. All image size is 1024 x 1024. 60 images are used for testing.

B. Implementation details

The proposed model is implemented on PyTorch framework. For training, the amount of data is increased through data augmentation. Specifically, 2,880 images for training were randomly cropped by 64 x 64. Also, these are rotated vertically and horizontally with a probability of 0.5. Our model requires the batch size to be a factor of 16 using Adam optimizer [15] with $\beta_1 = 0.9, \beta_2 = 0.999$. We train 100K iterations. Learning rate is 10^{-4} and retained throughout training. For evaluation, we measure Peak Signal-to-Noise Ratio (PSNR) in the YUV space and report Y-PSNR values.

Table **I**. PSNR performance comparisons

Method	perspective multi-view dataset (Synthetic)		MPEG multi-view dataset (Real)	
	Tar QP 37 Ref QP 32	Tar QP 42 Ref QP 37	Tar QP 37 Ref QP 32	Tar QP 42 Ref QP 37
Target (HEVC [1])	30.24	26.92	35.99	33.61
CrossNet [8]	30.42	27.17	36.09	33.74
Ours	**30.77**	**27.87**	**36.34**	**34.23**

C. Quantitative evaluation

We measure the PSNR between the ground-truth image and the compressed target image before applying each model for comparison, which is denoted by Target (HEVC) in Table 1. The PSNR values are 30. 24dB and 26.92dB when the target multi-view images in the synthetic dataset are coded with QP37 and QP42, respectively. Secondly, we measure the PSNR between the enhanced output through each network and the ground truth image. We compare the proposed model and CrossNet [8] which is a reference frame-based super-resolution model. For a fair comparison, CrossNet is trained on the same data. The results are shown in Table **I**. On average, our method achieves improved PSNR values approximately of 0.35-0.95dB. It is observed that the quality of the target multi-view images are significantly improved in terms of the average PSNR values. In addition, our method outperforms CrossNet by 0.25-0.7dB. The performance is worse than that of the proposed method.

D. Qualitative evaluation

We compare visual quality of the compared methods in Fig. 2 for qualitative evaluation. Fig. 2 shows the results of each model to enhance one frame in the test dataset. The first row shows the synthetic dataset result. The second row shows the test result using the dataset of camera captured scenes. Qualitative evaluation also shows the visual improvement on the target images.

V. CONCLUSION

In this paper, we proposed a novel model for multi-view image quality enhancement. We developed this model to apply on the distorted data due to compression as post-processing. To enhance the quality of the target image with relatively lower quality, it was proposed to use the multi-view images as references with relatively higher quality. The pipeline of MVIQE was full-convolutional, containing disparity estimation, backward warping, image reconstruction respectively. As a result of the experiment, it was confirmed that our proposed method has gain on the multi-view image enhancement. We will extend the proposed MVIQE to other low-level vision tasks in video frame quality enhancement in the future work.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00920, Development of Ultra High Resolution Unstructured Plenoptic Video Storage/Compression/Streaming Technology for Medium to Large Space). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1A2C4002052).

REFERENCES

[1] Sullivan, Gary J., et al. "Overview of the high-efficiency video coding (HEVC) standard." IEEE Transactions on circuits and systems for video technology 22.12 (2012): 1649-1668.

[2] J. Guo and H. Chao. Building dual-domain representations for compression artifacts reduction. In European Conference on Computer Vision(ECCV), pages 628-644, 2016.

[3] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Be-yond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing, 26(7):3142-3155, 2017.

[4] Kim, Hee-Jae, Je-Won Kang, and Byung-Uk Lee. "Super-resolution of multi-view ERP 360-degree images with two-stage disparity refinement." 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2020.

[5] Kim, Hee-Jae, Je-Won Kang, and Byung-Uk Lee. "360° Image Reference-Based Super-Resolution Using Latitude-Aware Convolution Learned From Synthetic to Real." IEEE Access 9 (2021): 155924-155935.

[6] Dong, Chao, et al. "Image super-resolution using deep convolutional networks." IEEE transactions on pattern analysis and machine intelligence 38.2 (2015): 295-307.

[7] Lim, Bee, et al. "Enhanced deep residual networks for single image super-resolution." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2017.

[8] ZHENG, Haitian, et al. Crossnet: An end-to-end reference-based super-resolution network using cross-scale warping. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018. p. 88-104.

[9] YANG, Fuzhi, et al. Learning texture transformer network for image super-resolution. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020. p. 5791-5800.

[10] Li, Zhengqi, et al. "Neural scene flow fields for space-time view synthesis of dynamic scenes." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[11] Niklaus, Simon, Ping Hu, and Jiawen Chen. "Splatting-based Synthesis for Video Frame Interpolation." arXiv preprint arXiv:2201.10075 (2022).

[12] SUN, Deqing, et al. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. p. 8934-8943.

[13] Bruhn, A., Weickert, J., Schn¨orr, C.: Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. IJCV 61(3) (2005) 211–231

[14] Jung, J., and B. Kroon. "Common test conditions for MPEG immersive video." 131th MPEG meeting of ISO/IEC JTC1/SC29/WG11. No. 19484. 2020.

[15] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[16] Nayoung Kim and Je-Won Kang, "Dynamic Motion Estimation and Evolution Video Prediction Network", IEEE Transactions on Multimedia, vol.23, pp 3986 – 3998, 2021

[17] Y. Chen, X. Zhao, L. Zhang, and J.-W. Kang, "Multiview and 3D Video Compression Using Neighboring Block based Disparity Vectors," IEEE Transaction on Multimedia, vol.18, no. 4, pp.576-589, Apr. 2016.

[18] Nayoung Kim, Seong Jong Ha, and Je-Won Kang, "Video Question Answering Using Language-Guided Deep Compressed-Domain Video Feature," International Conference on Computer Vision (ICCV), 2021

[19] Yu-Jin Ham, Chaehwa Yoo, and Je-Won Kang, "Training compression artifacts reduction network with domain adaptation," Proceedings Volume 11842, Applications of Digital Image Processing XLIV, 2021