Super-resolution of Multi-view ERP 360-Degree Images with Two-Stage Disparity Refinement

Hee-Jae Kim*,Je-Won Kang*, and Byung-Uk Lee*
* Department of Electronic and Electrical Engineering, Ewha W. University, Seoul 03760, Korea
E-mail: coolhj37@gmail.com, jewonk@ewha.ac.kr, bulee@ewha.ac.kr

***Abstract*—In this paper, we propose a novel super-resolution (SR) technique for multi-view 360-degree images in equi-rectangular projection (ERP) format. To the best of our knowledge, the proposed algorithm is the first study of multi-view 360-degree images in ERP. In multi-view SR (MV-SR), it is important to fuse the knowledge of features at different viewpoints, but the task is hardly achieved using a conventional CNN because conventional convolution is shift invariant. Thus, to solve the problem, we take a coarse-to-fine approach to exploit the correlation among multi-views in an ERP domain. First, we conduct depth-based warping on reference ERP to synthesize the image with the same viewpoint of the target low-resolution (LR) ERP. The non-linear distortion between the two ERP images can be remarkably reduced after the proposed warping. Second, we employ a flow estimator to refine the remaining flow between the warped reference image and the LR image. Our CNN architecture generates the SR at the end of the network by combining the features of LR-ERP and the warped reference ERP. It is demonstrated with experimental results that the proposed algorithm provides significantly improved quality of multi-view 360-degree images for SR as compared to the state-of-the-art in MV-SR.**

I. Introduction

Recently, 360-degree images are becoming popular. In virtual reality (VR), multi-view 360-degree cameras at different positions and their associated depth information allow more immersive experience to users [1]. The multi-view images require high-resolution (HR) representation to cover a wide range of spatial information. To process such large data, multi-view super-resolution (MV-SR) can be used for enhancing the experience in VR. For instance, a low-resolution (LR) of an image is transmitted in an encoder, and it can be rendered with the original HR in a decoder. Correspondence cues among adjacent views can be used for SR of the current view.

Equi-rectangular projection (ERP) is used as a defacto standard [6] to project a spherical view onto a 2D rectangular coordinate to form a 360-degree image. Different from typical plane projection, ERP sequences are distorted when projected to a 2D plane. Furthermore, the amount of distortion varies with the location of the camera in case of multi-view ERP sequence. Most of the previous work have focused on super-resolving stereo images and light field images using Convolutional Neural Network (CNN). In [2], [3], the CNNs are used for learning parallax between the left and the right images in stereo. In [4], [5], the CNNs are applied to exploit high similarity in sub-aperture images. However, few research on SR of multi-view 360-degree images in an ERP have been published. It is a challenging problem to learn a process to warp to a different viewpoint because of a curved non-linear epipolar line in ERP.

In this paper, we propose a novel SR technique for multi-view 360-degree images in ERP. In MV-SR, it is important to fuse the knowledge of features at different viewpoints to take more contexture information. To this aim, we take a coarse-to-fine approach to exploit the correlation among multi-views. First, we warp a reference ERP using the depth to synthesize an image with the same viewpoint of the LR-ERP. The non-linear distortion between the two ERP images can be remarkably eliminated after the proposed warping. Second, we use a flow estimator to refine the remaining flow between the warped reference image and the LR image. Our CNN architecture generates the SR at the end of the network by combining the features of LR-ERP and the warped reference ERP. It is demonstrated with experimental results that the proposed algorithm provides significantly improved quality of multi-view 360-degree images for SR.

The paper is organized as follows. We describe the proposed algorithm in Section III, the experimental results in Section V, and the remarks in Section VI.

II. Related Work

SR has been an interesting and important field of study in image processing and computer vision. Many traditional interpolation or dictionary-based sparse representation SR methods are superseded by learning based algorithms after wide spread use of deep learning. SRCNN [7] is the first CNN based single image SR (SISR) network announced by Dong et al. Further improvements are achieved by very deep SR (VDSR) [9] with deep layers and residual learning, efficient sub-pixel CNN(ESPCN) [8] which learns upscaling LR images, and enhanced deep SR (EDSR) [10] which optimizes the network with cascaded residual blocks and combine multiscale modules.

A. Multi-view Super-Resolution

MV-SR takes advantage of utilizing details in high-resolution image from different views. MV-SR can be divided into two categories depending on whether the reference image is adjacent to the LR image in a temporal domain or spatial domain. While video SR takes temporally adjacent frame(s) as a reference, MV-SR refers to an image from neighboring view point. MV-SR has been mainly done on light-field images

which consists of many sub-aperture images from different angular positions or stereo images which has left-and right-view. MV-SR based on stereo image is studied by Wang et al [2] and Jeon et al [3]. Wang learns parallax attention maps from the stereo image pair. Jeon's algorithm also learns parallax prior with the luminance and the chrominance learning models and then they are combined. However, these methods are not directly applicable on ERP dataset because multi-view ERP images cannot be aligned in principle. Other MV-SR are proposed to apply on the light-field dataset. Zhang et al [4] separate images in light field data depending on angular direction and feed them into residual network to raise angular resolution. Fan et al. [5] applies VDSR to individual images and then search similar patches from nearby angular positions and combine them in the network. Unfortunately, these methods either cannot be employed for upscaling an ERP sequence. Compared to very short baselines between light field images, non-linear and larger flows exist between multi-view ERP images.

III. Proposed Algorithm

A. Overview of Proposed Algorithm

The proposed method aims to estimate an HR image $\mathcal{V}_o^{SR}$ in the target view given an LR image $\mathcal{V}_o^L$ and an HR image $\mathcal{V}_r^H$ in the adjacent view r as a reference image. In the problem, $\mathcal{V}_o^L$ represents a downscaled image of the ground-truth HR image $\mathcal{V}_o^H$, and therefore it is important in MV-SR to provide an accurate reference image $\mathcal{V}_{r\to o}^H$ that is warped from r to the target view. We use a coarse-to-fine warping process because the estimation of the disparity is hardly achieved in an ERP images with a single attempt from the CNN trained using perspective projection imagery. We first use the depth-based warping procedure to capture non-linear distortion in adjacent views to generate a reference image. However, the warping becomes inaccurate due to occlusions and noise in the depth information. Therefore, the approximation is not enough. We use a flow estimator to correct the remaining discrepancy. The flow-based warping and the MV-SR are implemented in a CNN encoder-decoder architecture. In an encoder, a trained flow is applied to a feature map to correct for remaining discrepancy through convolution layers in the feature domain, and the SR is generated in a decoder. The overview of the proposed algorithm is shown in Fig 1.

B. Depth-based Warping in ERP

Disparity $d(p)$ refers to a disparity vector at a pixel p between two pixel domains. In case of warping from reference view r to target view o, the depth data from view r and the camera positions of both views in the 3D coordinates are required. First we map all the points of the reference view to the 3D world coordinate, using depth map and camera parameters. Then, we remap the pixels to the corresponding coordinate in the target by recalculating the latitude and longitude in the camera position of the view o. Based on the depth information, the reference view $\mathcal{V}_r^H$ is warped to the target view. Let us define $\mathcal{V}_{r\to\tilde{o}}^H(p)$ as a sample value at pixel p, warped from the reference. Then, we have

$$\mathcal{V}_{r\to\tilde{o}}^H(p) = \mathcal{V}_{r\to\tilde{o}}^H(p_r + d_r(p_r)) = \mathcal{V}_r^H(p_r), \tag{1}$$

where p corresponds to p_r in $\mathcal{V}_r^H$.

Disparity d_r has a sub-pixel precision, therefore the warped pixel tends to be misaligned from the integer-grid in the pixel-coordinates. There can be two different kinds of problems when a pixel is mapped to the nearest integer-grid: mutiple mapping and occlusion. In one case, if there are more than one correspondence to p, the samples are overlapped. We select a pixel in the foreground with smaller depth. In the other case, there can be none of the correspondence by occlusion, which incurs a hole during the warping. We fill out the holes by using the flow-based warping in sequel.

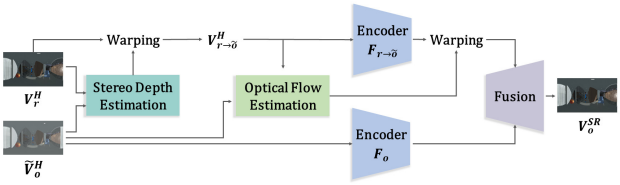

Fig. 1. Overview of our proposed MV-SR method with two-stage disparity refinement module. First, we estimate the stereo depth to warp the reference image toward the target viewpoint. Then we align the reference image with high accuracy by estimating remaining flows between the warped reference image and the target image. Finally, we conduct RefSR on target image with refined reference imgae.

C. Flow-based Warping

We apply flow-based warping using deep learning model. The warping is performed in a CNN encoder as shown in Fig.1. The encoder produces a set of feature maps $\mathcal{F} \in \mathbb{R}^{f_h \times f_w \times f_c}$ through convolution layers wheref_h, f_w, and f_c represent the height, the width, and the number of the feature maps.

In feature extraction, $\tilde{V}_o^H$ as an up-sampled LR-ERP through bi-linear interpolation and $V_{r\to\tilde{o}}^H$ as the depth-based warped HR-ERP go through the same network, independently. Both inputs generate the feature maps $F_{r\to\tilde{o}}$ and F_o, respectively. Unlike the conventional algorithms, the proposed network computes a flow vector f and apply the refitment in the feature domain. The refinement is represented as follows:

$$\mathcal{F}_{r\to\tilde{o}\to o}(p) = \mathcal{F}_{r\to\tilde{o}}(p + f), \tag{2}$$

where p is the coordinates in the feature map. We adopt FlowNet-Simple [11] for estimating the flow.

D. Multi-View Super-Resolution

In a decoder, the extracted features are concatenated after the refitment, and they are up-scaled using three deconvolution layers and three convolution layers for fusion. At the end of the network, all the features are fused for generating

the SR V_o^{SR}. While the flow estimator is pretrained FlowNet-Simple [11], the feature extraction and SR output networks are trained with multi-view sequence.

IV. Dataset

A. Classroom

In experiments, we use a "Classroom" dataset from MPEG-I standard activity. The dataset has multi-view 360-degree images from fifteen views as shown in Fig. 2. Around a center view (v_0), there are six cameras ($v_1 \sim v_6$) in the inner circle of the radius 6 cm. The group is named G_{IC}. A six cameras ($v_9 \sim v_{14}$) are located in the outer circle with the radius 11.2 cm. The group is named G_{OC}. The cameras are equally spaced in each group of the circles. In addition, there are top view (v_8) and bottom view (v_7) named G_{TB}. The images have multi-view video plus associated depth data. The image and depth map have the same resolution of 4096 $\times$ 2048. There are 120 images in each view. It is also noted that the "Classroom" is formatted with a full ERP, which covers 360 $\times$ 180 degrees of field of view.

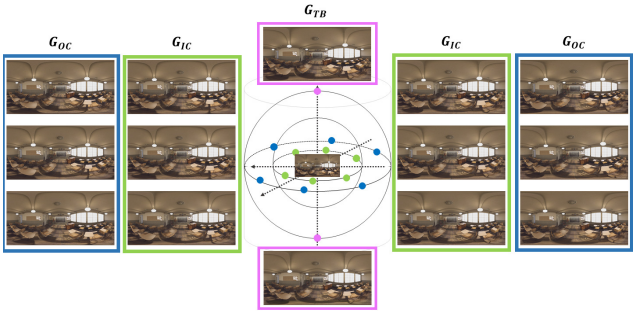

Fig. 2. View distribution of Classroom. G_{IC} includes $v_1 \sim v_6$, G_{OC} includes $v_9 \sim v_{14}$ and G_{TB} includes v_7 and v_8

B. INDOOR360

INDOOR360 dataset has 20 synthetic multi-view ERP sequences, including human avatars, animals, and other 3D figures as foreground objects with different textures. The objects are rendered in various indoor backgrounds such as in schools and houses. The scenes are captured from four different viewpoints where each of three cameras $p_2 \sim p_4$ is placed 25cm away in the x, y, and z directions around the center viewpoint p_1. The dataset is also formatted with a full ERP. Unlike "Classroom", depth maps in INDOOR360 are generated with 360 SD-Net [12] as the state-of-the-art stereo depth estimation network. 360 SD-Net [12] estimates the disparity between any two views.

V. Experimental results

In this section, both quantitative and qualitative comparisons are presented to evaluate the performance of proposed MV-SR. Moreover, we investigate the effectiveness of depth-based warping layer depending on the various range of disparity. All the experiments are performing $\times$4 SR, which enhance the LR-ERP image at v_0 in Classroom and p_1 in INDOOR360.

A. Performance Evaluation

We compare the performance between the proposed algorithm and other SR schemes using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). The SISR methods in the comparisons are SRCNN [7], VDSR [9] and EDSR [10]. We also compared the performance with PASSRNet [2] as the state-of-the-art MV-SR. In case of MV-SR, the HR-ERP at v_7 in Classroom and p_4 in INDOOR360 are utilized as the reference frame.

Table I shows the performance comparisons. Owing to the reference image, the proposed algorithm achieves the best performance in Classroom and the comparable performance with PASSRNET [2] in INDOOR360 on average. The Classroom dataset includes the ground-truth depth data while INDOOR360 has only the estimated one. For this reason, the depth-based disparity refinement layer provides more accurate warping process on Classroom than INDOOR360. Nevertheless, the proposed algorithm outperforms SISR methods in INDOOR360.

TABLE I
Comparison of quantitative results (PSNR / SSIM) of ($\times$4 SR) between SISR methods and Proposed on Classroom and INDOOR360.

Algorithm		Classroom	INDOOR360
SISR	Bicubic	27.34 / 0.90	25.86 / 0.80
	SRCNN [7]	28.73 / 0.92	26.59 / 0.83
	VDSR [9]	28.88 / 0.92	26.65 / 0.83
	EDSR [10]	29.75 / 0.93	27.37 / 0.85
MV-SR	PASSRNet [2]	30.02 / 0.94	**27.56** / 0.85
	Proposed	**31.85 / 0.95**	27.48 / **0.86**

The visual quality of SR images is compared in Fig.3 using both the datasets. Since the similarity between the target and the reference image is a key to the success of MV-SR, we focused on the process of exploiting the correlation among multi-views with depth-based warping. As a result, our method shows better visual quality with clear textures. The PASSRNet [2] and CrossNet [13] also provide fair performance as compared to SISR by benefiting from the HR reference image. However, the proposed algorithm outperforms them when the reference image is well-aligned and the textures can be efficiently super-resolved. For example, CrossNet [13] is incapable of handling large discrepancy between the images in the second example and the misaligned HR details are remained in results. However, our method effectively eliminates the difference and shows fine textures in the result.

B. Ablation Study

We proposed the depth-based warping layer to effectively overcome the large disparity between the multi-views for RefSR. In order to benchmark the proposed depth-based disparity refinement layer depending on the range of disparity, we conduct an ablation study by removing the disparity refinement layer. We compare the results at four different disparity levels. In Table II, we turned off the depth-based warping process denoted by "Proposed-D" and compare the performance with the

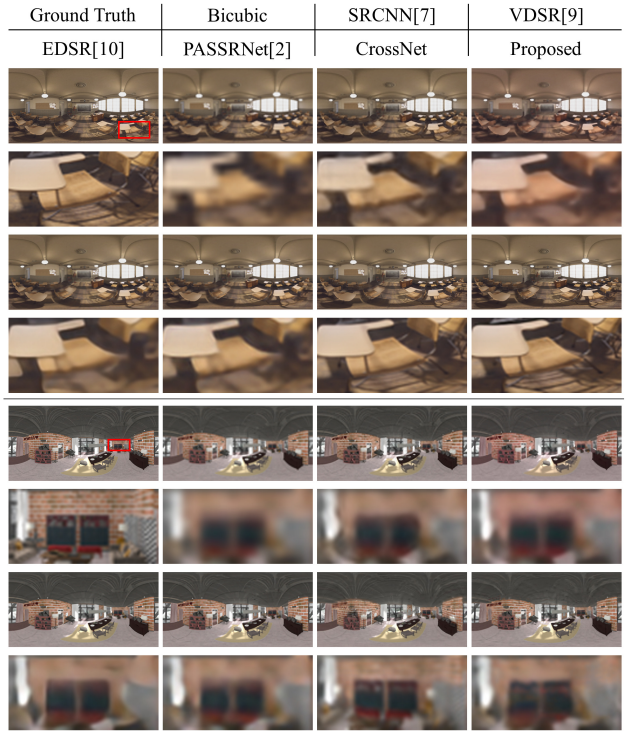

Fig. 3. Qualitative comparison among different SR methods on Classroom in the first four rows and INDOOR360 in the next four rows.

original one denoted by "Proposed". Given that the references in the G_{IC}, G_{OC} and G_{TB} are located 6 cm, 11.2 cm, 6 cm away from the center in Classroom and INDOOR360 is placed 25cm away in INDOOR, they represent the varying disparity levels.

Overall, "Proposed" outperforms "Proposed-D" in Table II. We observe there were more improvements when the disparity becomes larger, which reveals that the coarse-to-fine approach effectively works for refining the disparity. In other words, the "Proposed-D" on INDOOR360 is not enough to compensate the large difference between the multi-views. In this case, the LR could not avoid to refer the misaligned HR information.

TABLE II
ABLATION STUDY ON PROPOSED $\times 4$ MV-SR WITH CLASSROOM AND INDOOR360 DEPENDING ON THE RANGE OF THE DISPARITY BETWEEN MULTI-VIEWS. "PROPOSED-D" DENOTES THE MV-SR NETWORK WITHOUT DEPTH-BASED WARPING PROCESS. PSNR / SSIM COMPARISON BETWEEN "PROPOSED" AND "PROPOSED-D".

Group	G_{IC}	G_{OC}	G_{TB}	INDOOR360
Proposed	31.80 / 0.95	30.59 / 0.93	33.17 / 0.97	27.48 / 0.86
Proposed-D	29.20 / 0.94	27.98 / 0.92	31.83 / 0.96	25.07 / 0.79

VI. CONCLUSIONS

A SR technique for multi-view 360-degree images was firstly proposed in this paper. In the proposed technique, we used a coarse-to-fine approach to exploit the correlation among multi-views. First, we warped a reference ERP with the depth to the same viewpoint of the LR-ERP. Second, we used a flow estimator to refine the remaining flow between the warped reference image and LR image. The SR was generated by combining the features of LR-ERP and the warped reference ERP. It was demonstrated that the proposed algorithm outperforms the state-of-the-arts in MV-SR.

ACKNOWLEDGMENT

This work was supported by Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00765 Development of Compression and Transmission Technologies for Ultra High Quality Immersive Videos Supporting 6DoF) and the National Research Foundation of Korea (NRF) (No.NRF-2019R1C1C1010249)

REFERENCES

[1] G. Simon, X. Corbillon, F. Simone, and P. Frossard, "Dynamic adaptive streaming for multi-viewpoint omnidirectional videos," in *MMSys Proceedings of the 9th ACM Multimedia Systems Conference,* 2018.

[2] L. Wang, Y. Wang, Z. Liang, Z. Lin, J. Yang, W. An and Y. Guo, "Learning parallax attention for stereo image super-resolution," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2019, pp.12250-12259.

[3] D. S. Jeon, S. H. Baek, I. Choi, and M.H. Kim, "Enhancing the spatial resolution of stereo images using a parallax prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2018, pp. 1721-1730.

[4] S. Zhang, Y. Lin, and H. Sheng, "Residual networks for light field image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2019, pp. 11046-11055.

[5] H. Fan, D. Liu, Z. Xiong, and F. Wu, "Two-stage convolutional neural network for light field super-resolution," in *IEEE International Conference on Image Processing,* 2017, pp. 1146-1171.

[6] Y. Ye, E. Alshina, and J. Boyce, ""Algorithm description of projection format conversion and video quality metrics in 360lib," in *Joint Video Exploration Team of ITU-T SG, 16,* 2017.

[7] C. Dong, C. C. Loy, K. He, "Image super-resolution using deep convolutional networks," in *IEEE transactions on pattern analysis and machine intelligence, 38(2),*, 2015, pp. 295-307.

[8] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, ... and Z. Wang, "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,* 2016, pp. 1874-1883.

[9] J. Kim, J. K. Lee, and Mu Lee, K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2016, pp. 1646-4654.

[10] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops,*, 2017, pp. 136-144.

[11] A. Dosovitskiy, P. Fisher, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, ... and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *IEEE international conference on computer vision,* 2015, pp. 2758-2766.

[12] N. H. Wang, B. Solarte, Y. H. Tsai, W. C. Chiu, and M. Sun, "360SD-Net: 360° Stereo depth estimation with learnable cost volume," in *arXiv preprint arXiv:1911.04460 2,* 2019.

[13] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang, "Crossnet: An end-to-end reference-based super resolution network using cross-scale warping," in *In Proceedings of the European Conference on Computer Vision,* 2018, pp.88-104.