Received May 22, 2022, accepted May 27, 2022, date of publication June 2, 2022, date of current version June 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3179834

Light Field Image Super-Resolution Based on Multilevel Structures

HWA-JONG PARK[1], JUNHO SHIN[2], HYUNSEOP KIM[2], AND YEONG JUN KOH[2]

[1]Intekplus Company, Daejeon 34026, South Korea
[2]Department of Computer Science and Engineering, Chungnam National University, Daejeon 34134, South Korea

Corresponding author: Yeong Jun Koh (yjkoh@cnu.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. NRF-2021R1A4A1031864 and No. NRF-2022R1I1A3069113) and in part by Institute of Information & communications Technology Planning & evaluation (IITP) grant funded by the Korea government (MSIT) (No.2020-0-01441, Artificial Intelligence Convergence Research Center (Chungnam National University)).

ABSTRACT Light field (LF) images suffer from low spatial resolution due to the trade-off between angular and spatial resolutions. Thus, spatial super-resolution (SR) of LF images is an essential task to obtain high-quality LF images. However, the existing SR networks still have limitations, since they exploit only single-level features to use sub-pixel information in LF images. In this paper, we proposed a light field super-resolution (LFSR) network to effectively improve the spatial resolution of light field images. The proposed network takes one target image and its 8-neighboring images for references. We construct multi-level structures for the proposed network to effectively estimate and mix sub-pixel information in reference images. The proposed network is composed of a feature extractor, a feature warping module, a feature mixing module, and a upscaling module. The feature extractor provides multi-level features for SR and offsets to the feature warping module to obtain aligned features for multiple reference images. The feature mixing module mixes multiple aligned features based on the similarity between the target and reference images to obtain multi-level mixed features. Finally, the upscaling module generates a high-resolution residual image using the multi-level mixed features. Experimental results demonstrate the proposed network outperforms the state-of-the-art methods on various light field datasets. The pre-trained model and source codes are available at https://github.com/Hwa-Jong/LF_MLS.

INDEX TERMS Light field, super-resolution, light field super-resolution, convolutional neural network, multi-level feature, deformable convolution.

I. INTRODUCTION

Light field cameras record not only spatial information but also angular information by inserting a micro-lens array between the main lens and the image sensor [1]. From recorded spatial and angular data, multi-view images of a scene can be reconstructed. These light field images have been used in many computer vision tasks, such as saliency detection [2], [3], depth sensing [4], [5], de-occlusion [6]–[8]. However, light field images have low spatial resolution due to the trade-off between angular and spatial resolutions. Low spatial resolution images lead to performance degradation of computer vision applications. Therefore, light field image super-resolution (LFSR) is required to improve the performance of the applications. To solve this problem, we propose a LFSR network to achieve spatial super-resolution based on multi-level structures.

Since light field images are highly correlated with each other, sub-pixel information can be estimated using adjacent view images. Different from traditional single image super-resolution (SISR), LFSR can generate high-resolution images using sub-pixel information estimated from other light field images. Recently, with the release of large LFSR datasets [7], [9]–[13], many deep learning networks [14]–[18] have been developed based on convolutional neural networks (CNNs), which uses multiple view images as references. Figure 1 shows some approaches in the existing LFSR methods using multiple view images. All LF images are used to generate all super-resolved LF images (all-to-all) [15], [19] or each super-resolved view image (all-to-one) [18].

The associate editor coordinating the review of this manuscript and approving it for publication was Junhua Li.

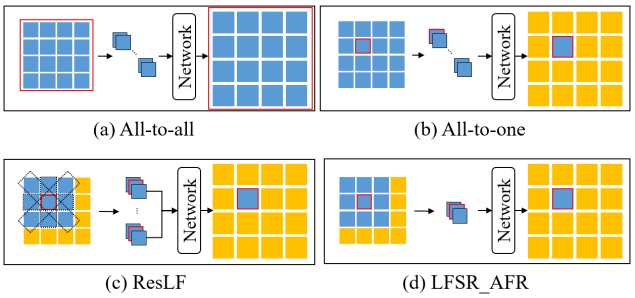

FIGURE 1. Examples of approaches in SR of 4 × 4 LF images: (a) all-to-all [15], [19], (b) all-to-one [18], (c) ResLF [14], and (d) LFSR_AFR [16]. Red border boxes represent target images, which are super-resolved, and blue boxes are reference images.

Also, ResLF [14] uses several sets of LF images in various directions. These approaches require the fixed angular resolution of LF images to obtain super-resolved LF images. On the other hand, the approach in LFSR_AFR [16], which uses 8-neighboring view images, can provide super-resolved LF images regardless of angular resolutions. However, LFSR_AFR [16] still has limitation that it exploit only single-level features to increase the spatial resolution.

In this paper, we propose a light field super resolution network, which enhance the spatial resolution of LF images, based on multi-level structures. The proposed network consists of a feature extractor, a feature warping module, a feature mixing module, and a upscaling module. The proposed network takes one target view image and its 8-neighboring view images for references. In the feature extractor, the proposed network extracts low-level features, as well as high-level features for each view image. Then, the feature warping module warps features of reference images to the target image using deformable convolution to obtain aligned features. Next, the feature mixing module yields multi-level mixed features by combining the multiple aligned features of reference images based on the similarity between the reference images and the target image. Finally, the multi-level mixed features are used to produces a super-resolved residual image in the upscaling module. Experimental results demonstrate that the proposed LFSR network outperforms the state-of-the-arts on various LF datasets without increasing the number of parameters.

II. RELATED WORK

Image SR aims to generate a high-resolution image from its low-resolution image. However, image SR is a classical ill-posed problem [20]. To solve this problem, many CNN-based SR networks have been developed to improve the SR performance. SR can be categorized into single image super-resolution (SISR), video super-resolution (VSR), and light filed super-resolution (LFSR).

A. SISR

The goal of SISR is to generate a high-resolution image from only one low-resolution image. SRCNN [21] is the first SISR network, which is composed of only three convolution layers. Also, VDSR [22] including 20 convolution layers was developed. After that, many CNN-based SISR networks have been developed, including attention-based [23], [24], and generative adversarial networks-based [25], [26] methods.

B. VSR

VSR attempts to increase the spatial resolution of frames based on temporally adjacent frames. In terms of exploiting sub-pixel information of reference frames through motions, VSR is similar to LFSR. DUF [27] generated high-resolution frames using dynamic upsampling filters. Tian *et al*. [28] adopted deformable convolution [29] for VSR. EDVR [30] reconstructed high-resolution frames using a deformable alignment module and temporal and spatial attention fusion modules.

C. LFSR

LFSR aims at generating high-resolution LF images from low-resolution LF images. To learn mapping between low and high-resolution light field images based on data-driven method, Yoon *et al*. [31] proposed an early model for LFSR based on deep CNNs. Fan *et al*. [32] proposed two-stage CNNs for SISR and multi-patch fusions. Gul and Gunturk [19] proposed two networks for angular SR and spatial SR.Wang *et al*. [33] improved horizontally and vertically stacked images separately and combined the using stacked generalization. Zhang *et al*. [14] divided view images into four groups and stacked views in each group to use residual information between neighbor views. Yeung *et al*. [34] generated high-resolution LF images using the spatial-angular separable convolution. Ko *et al*. [16] proposed two networks to improve spatial and angular resolution based on the adaptive feature remixing. Jin *et al*. [18] proposed an all-to-one strategy for LFSR, which enforces the LF parallax structure in reconstructed LF images.

III. PROPOSED NETWORK

We adopt the 4D light field representation in [35]. Thus, the light field can be represented as:

$$\mathbf{L}(u, v, x, y) \in \mathbb{R}^3, \tag{1}$$

where $\mathbf{L}$ is in the color space, such as RGB space. Also, (u, v, x, y) are defined on the domain $\mathbb{N}_U \times \mathbb{N}_V \times \mathbb{N}_W \times \mathbb{N}_H$, where $\mathbb{N}_k \triangleq \{1, 2, \ldots, k\}$. Also, (u, v) denotes an angular coordinate and (x, y) denotes a spatial coordinate. Thus, $\mathbf{L}$ has $U \times V$ light field images of $H \times W$ spatial resolution. Let $I_{\mathbf{u}}$ denote a view image at an angular coordinate $\mathbf{u} = (u, v)$. The proposed network super-resolves each view image $I_{\mathbf{u}}$ to reconstruct higher-resolution light field

$$\mathbf{L}^{HR}(u, v, x, y) \in \mathbb{R}^3, \tag{2}$$

defined on $\mathbb{N}_U \times \mathbb{N}_V \times \mathbb{N}_{rW} \times \mathbb{N}_{rH}$ with a scale factor r.

A. NETWORK ARCHITECTURE

Figure 2 shows the overview of the proposed network. To increase the spatial resolution of a view image $I_{\mathbf{u}}$, the

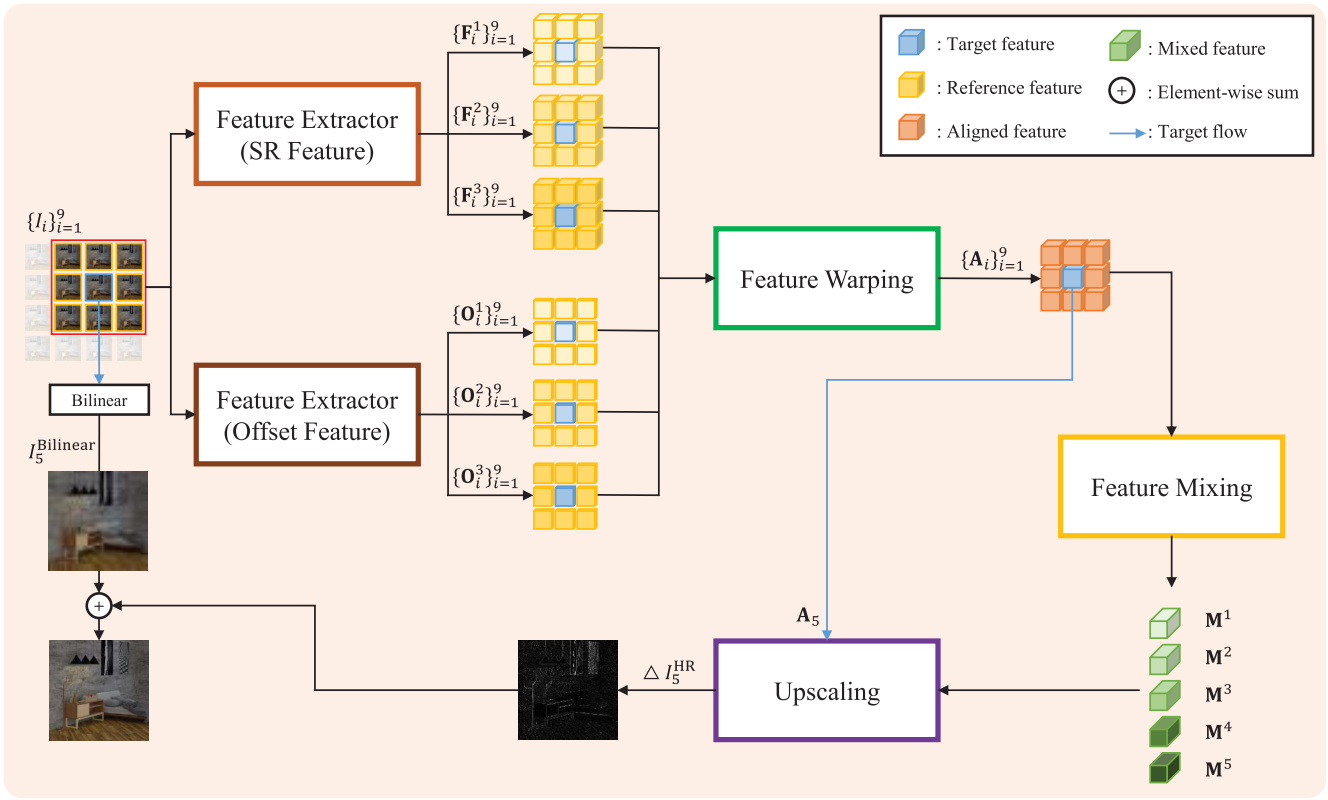

FIGURE 2. Architecture of the proposed network.

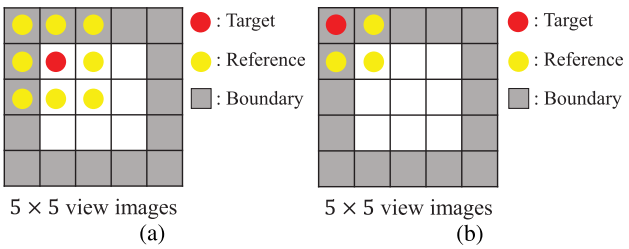

FIGURE 3. The different number of available reference images according to angular coordinates of the target image.

proposed network additionally takes 8-adjacent view images in the angular domain to use sub-pixel information. Let $\mathcal{I}_{\mathbf{u}} = \{I_i\}_{i=1}^{9}$ denote 3×3 view images centered on $I_{\mathbf{u}}$. They are indexed from top-left to bottom-right, and thus we refer to $I_5 = I_{\mathbf{u}}$ as a target image and $\{I_i\}_{i=1,i\neq 5}^{9}$ as reference images. As in Figure 3(a), a target image has 8 reference images when it is located in center of light field. On the other hand, some reference images are unavailable when a target image is located in the boundary of light field as in Figure 3(b). In this case, we use virtual images filled with zero value for unavailable reference images. To this end, given $\mathcal{I}_{\mathbf{u}}$, the proposed network estimates a super-resolved image $I_{\mathbf{u}}^{HR}$ through feature extractor, feature warping, feature mixing, and upscaling modules.

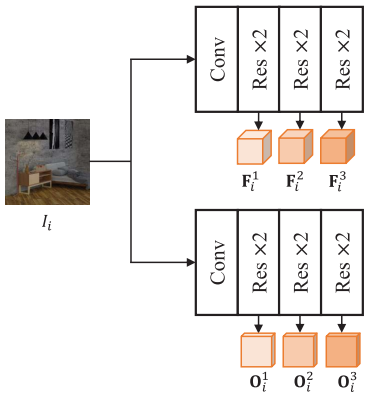

FIGURE 4. Structure of the feature extractor.

1) FEATURE EXTRACTOR

The feature extractor consists of a convolution layer and six residual blocks. It takes each view image and extracts a feature for every two residual blocks to construct multi-level features. Thus, 3-level features are obtained from the feature extractor. As in Figure 4, we use two feature extractors of the same structure to extract SR and offset features. To this end, for each I_i, multi-level SR features $\{\mathbf{F}_i^l \in \mathbb{R}^{H\times W\times C}\}_{l=1}^{3}$ and multi-level offset features $\{\mathbf{O}_i^l \in \mathbb{R}^{H\times W\times C/2}\}_{l=1}^{3}$, where $C = 32$, are obtained.

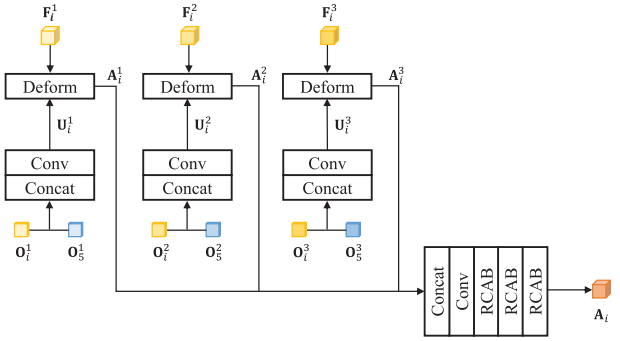

FIGURE 5. Structure of the feature warping module.

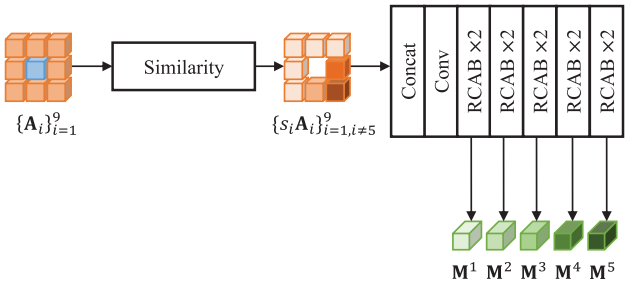

FIGURE 6. Structure of the feature mixing module.

2) FEATURE WARPING MODULE

Reference images in $\mathcal{I}_{\mathbf{u}}$ contain sub-pixel information for SR, but there are various sub-pixel shifts according to angular positions such as horizontal, vertical, and diagonal offsets. To use sub-pixel information effectively, SR features of reference images should be aligned to the target image I_5. The feature warping module estimates offsets between the target image I_5 and reference images to warp SR features of reference images to the target frame.

Figure 5 illustrates the detailed structure of the feature warping module. For each reference image I_i, $i \neq 5$, offset features $\mathbf{O}_i^l$ and $\mathbf{O}_5^l$ are concatenated, and the concatenated one is fed into a convolution layer to obtain an offset $U_i^l \in \mathbb{R}^{H\times W\times 2}$. Then, by employing the deformable convolution [36], $\mathbf{F}_i^l$ is warped to the target image to obtain an aligned SR feature $\mathbf{A}_i^l$. Here, U_i^l is used for the offset in the deformable convolution. This warping process is performed for all feature levels, and thus multi-level aligned features for I_i, $\{\mathbf{A}_i^l\}_{l=1}^3$, are obtained.

Warping results at lower levels tend to preserve detailed local motions, while those at higher levels contain global shifts between target and reference images. To explore these multi-level features effectively, the feature warping module combines them based on RCAB [23]. Specifically, multi-level aligned features are concatenated along the channel dimension, and then the concatenated feature sequentially passes through one convolution layer and three RCABs to form the combined aligned feature $\mathbf{A}_i$. Since RCAB is the channel attention block, it can generate more effective feature for SR from the concatenated feature. Unlike the reference images, the multi-level SR features for the target frame are combined without the deformable convolution, since the SR features for the target frame do not need the warping processing. To this end, the feature warping module provides the set of aligned features $\{\mathbf{A}_i\}_{i=1}^9$ for all view images.

3) FEATURE MIXING MODULE

Figure 6 illustrates the detailed structure of the feature mixing module. The feature mixing module combines the aligned features of the reference images to explore sub-pixel information for SR of the target image. However, some reference images have the low reliability to use sub-pixel information. For instance, as in Figure 3(b), reference images are filled zero images, when the target image is on the boundary of LF. To alleviate the impact of those dummy images, we compute similarity scores between the target image and reference images.

For each reference image I_i, $i \neq 5$, the feature mixing module compares aligned features $\mathbf{A}_5$ and $\mathbf{A}_i$ through point-wise dot product. Thus, the similarity score s_i is defined as

$$s_i = 8\frac{\exp(\mathrm{Tr}(\tilde{\mathbf{A}}_5 \times \tilde{\mathbf{A}}_i^T))}{\sum_{i=1,i\neq 5}^{9}\exp(\mathrm{Tr}(\tilde{\mathbf{A}}_5 \times \tilde{\mathbf{A}}_i^T))} \tag{3}$$

where $\mathrm{Tr}(\cdot)$ denotes a trace operation and $\tilde{\mathbf{A}}_i \in \mathbb{R}^{HW\times C}$ is the reshaped matrix of $\mathbf{A}_i$. Then, the similarity scores are used for weights of aligned features.

Given weighed aligned features of reference images, $\{s_i\mathbf{A}_i\}_{i=1,i\neq 5}^9$, the feature mixing module combines them using several RCABs. The weighed aligned features of reference images are concatenated, and then the concatenated feature sequentially passes through one convolution layer and ten RCABs to obtain mixed features. The feature mixing module extracts mixed features from every two RCABs. To this end, the feature mixing module produces multi-level mixed features $\{\mathbf{M}^l\}_{l=1}^5$.

4) UPSCALING MODULE

The upscaling module takes $\mathbf{A}_5$ and the multi-level mixed features $\{\mathbf{M}^l\}_{l=1}^5$ to increase the spatial resolution of the target image I_5. Figure 7 illustrates the architecture of the upscaling module. To consider the multi-level mixed features sequentially, we use several U-blocks, each of which is composed of one convolution layer and three residual blocks. Each U-block takes the mixed feature of each level and the output of the previous U-block as an input and sequentially processes it to extract the higher-level feature.

Also, we adopt the information pool [37] to combine multi-level outputs of U-blocks. The information pool analyzes outputs of the first to fifth U-blocks to extract a information feature P. The information feature P is used in the last four U-blocks as in Figure 7. U-blocks at the same level are linked with skip connections as done in U-net [38]. The upscaling module adds the output of the last U-block to $\mathbf{A}_5$,

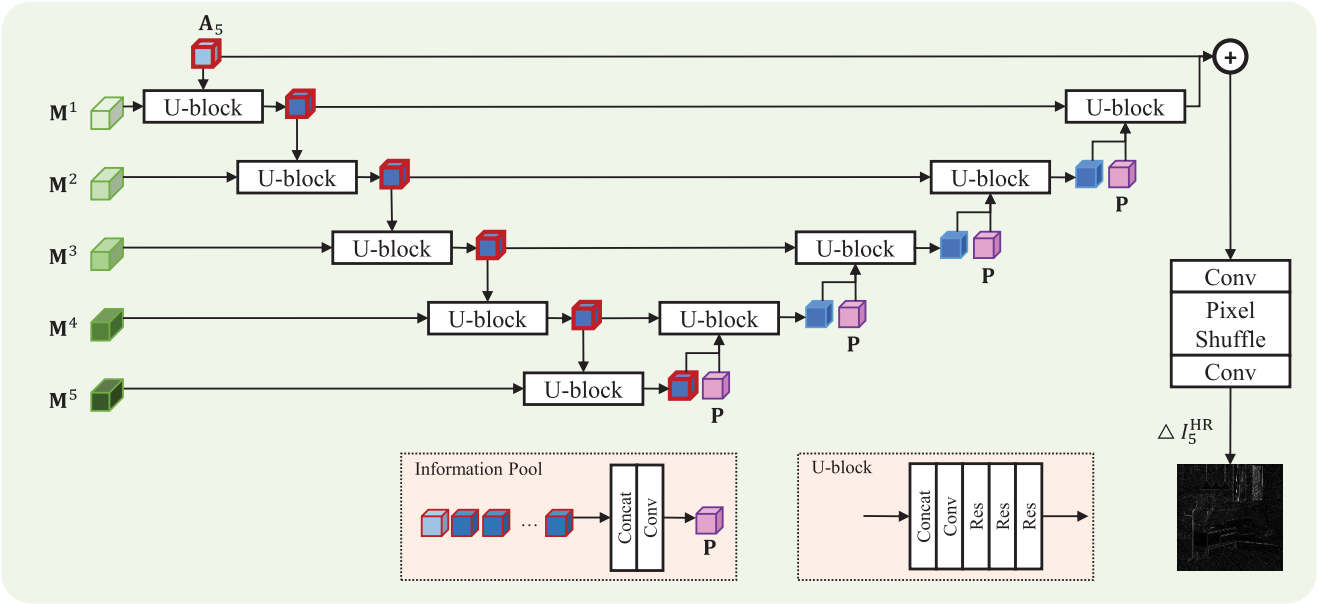

FIGURE 7. Structure of the upscaling module. Red bordered features are fed to information pool to generate the information feature.

TABLE 1. List of LF images in the test set.

Dataset (# scenes)	Scenes				
HCI (4)	*Buddha*	*Horses*	*Mona*	*Papillon*	
HCI2 (4)	*Bedroom*	*Bicycle*	*Boxes*	*Sideboard*	
EPFL (8)	*Flowers University*	*Friends 5 Paved Road*	*Fountain Pool Palais du Luxembourg*	*ISO Chart 12*	*Reeds*
Stanford (14)	*Bikes 1-2 People 1-2*	*Buildings 1-2 Reflective 1-2*	*Occlusions 1-2*	*General 1-2*	*Cars 1-2*
INRIA (5)	*Bee_1*	*Building*	*Hublais*	*MessyDesk*	*Sculpture*

and the pixel-shuffle layer [39] generates a super-resolved residual image $\triangle I_5^{HR}$, whose spatial resolution is $rH \times rW$. Finally, the super-resolved image for the target image I_5^{HR} is defined as

$$I_5^{HR} = I_5^{Bilinear} + \triangle I_5^{HR}, \tag{4}$$

where $I_5^{Bilinear}$ is a up-sampled target image using bilinear interpolation.

5) IMPLEMENTATION DETAILS

As done in [14]–[16], [33], we convert RGB color space into YCbCr color space and try to super=resolve only Y color space. Cb and Cr colors are up-sampled using bicubic interpolation. We use the L_1 loss function between the predicted super-resolved image and the ground-truth. We use the Adam [40] optimizer and the leakyReLU [41] with the slope of 0.2 for negative input as the activation function. The batch size is set 32. Also, the learning rate is initially set to 0.001 and decreased by a factor of 0.5 for every 50 epochs. We stop the training after 300 epochs.

IV. EXPERIMENTAL RESULTS

We first perform ablation studies to demonstrate the components in the proposed network. Next, we compare the proposed network with the state-of-the-arts networks, including [14], [16].

A. DATASET AND METRIC

For experiments, we use HCI [9], HCI2 [10], EPFL [12], Stanford [11], and INRIA [7]. Here, HCI and HCI2 are synthetic datasets, whereas EPFL, Stanford, and INRIA are real-world datasets. For the fair comparison, we use the same training set as [14], [16], which is composed of 246 LF images. Also, Table 1 lists test LF images used in our experiments. For each scene in the test set, 9×9 view images are used for the evaluation. For quantitative evaluation, we use

TABLE 2. Ablation studies on the proposed network.

Settings	#Params.	Datasets				
		HCI [9]	HCI2 [10]	EPFL [12]	Stanford [11]	INRIA [7]
w/o reference images	0.78M	39.04/0.981	34.69/0.964	36.99/0.977	37.90/0.984	34.89/0.976
w/o feature warping	1.45M	41.89/0.989	37.24/0.980	37.54/0.980	39.49/0.989	36.77/0.981
w/o feature mixing	1.45M	41.81/0.989	37.16/0.980	37.54/0.980	39.40/0.988	36.63/0.981
w/o multi-level structure	1.49M	41.97/0.989	37.29/0.980	37.54/0.980	39.48/0.989	36.83/0.981
Proposed	1.58M	**42.07/0.989**	**37.33/0.980**	**37.56/0.981**	**39.57/0.989**	**36.95/0.981**

TABLE 3. Comparison of the proposed network with the existing networks in terms of PSNR/SSIM scores for scale factor ×2 and for all view images. The best results are boldfaced, and the second best ones are underlined.

	Datasets				
	HCI [9]	HCI2 [10]	EPFL [12]	Stanford [11]	INRIA [7]
Bicubic	35.23/0.930	31.67/0.882	31.23/0.886	31.79/0.949	29.53/0.945
LFNet [33]	36.46/0.964	33.63/0.932	32.70/0.935	35.45/0.975	32.39/0.966
EDSR [42]	39.24/0.966	35.07/0.949	33.94/0.947	32.33/0.958	30.75/0.959
SOF-VSR [43]	39.12/0.959	34.75/0.932	34.61/0.934	32.15/0.956	30.49/0.957
ResLF [14]	41.09/0.988	36.45/0.979	35.48/0.973	38.88/0.987	34.69/0.980
LFSR_AFR [16]	42.06/**0.989**	37.27/**0.980**	37.21/0.977	39.44/0.988	35.62/0.975
Proposed	**42.07/0.989**	**37.33/0.980**	**37.56/0.981**	**39.57/0.989**	**36.95/0.981**

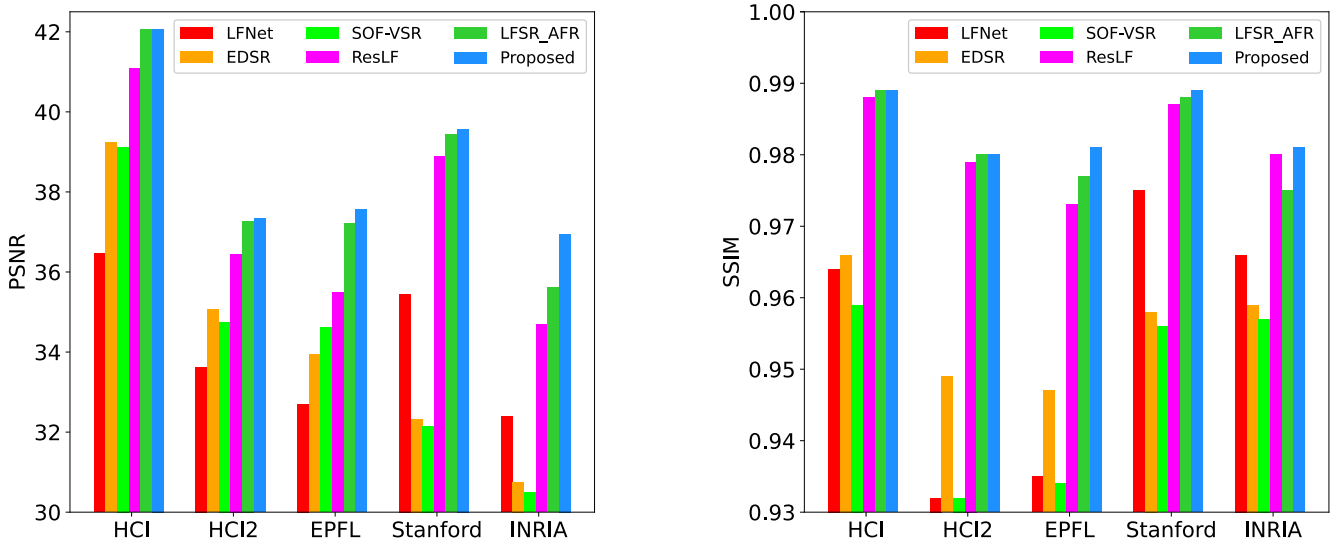

FIGURE 8. Bar plots of PSNR/SSMI scores for comparisons.

the PSNR/SSIM scores between the original high-resolution image and the super-resolved image.

B. ABLATION STUDIES

We perform ablation studies to demonstrate the effectiveness of the components in the proposed algorithm. Table 2 shows the comparative results between the proposed network and its variations.

1) WITHOUT REFERENCE IMAGES

We do not use reference images to enhance the spatial resolution of the target image. In other words, the network takes only the target image as an input. Also, in this setting, the feature warping module and the feature mixing module are excluded from the proposed network, since there are no reference features. As in Table 2, without sub-pixel information in reference images, the network provides unreliable SR results.

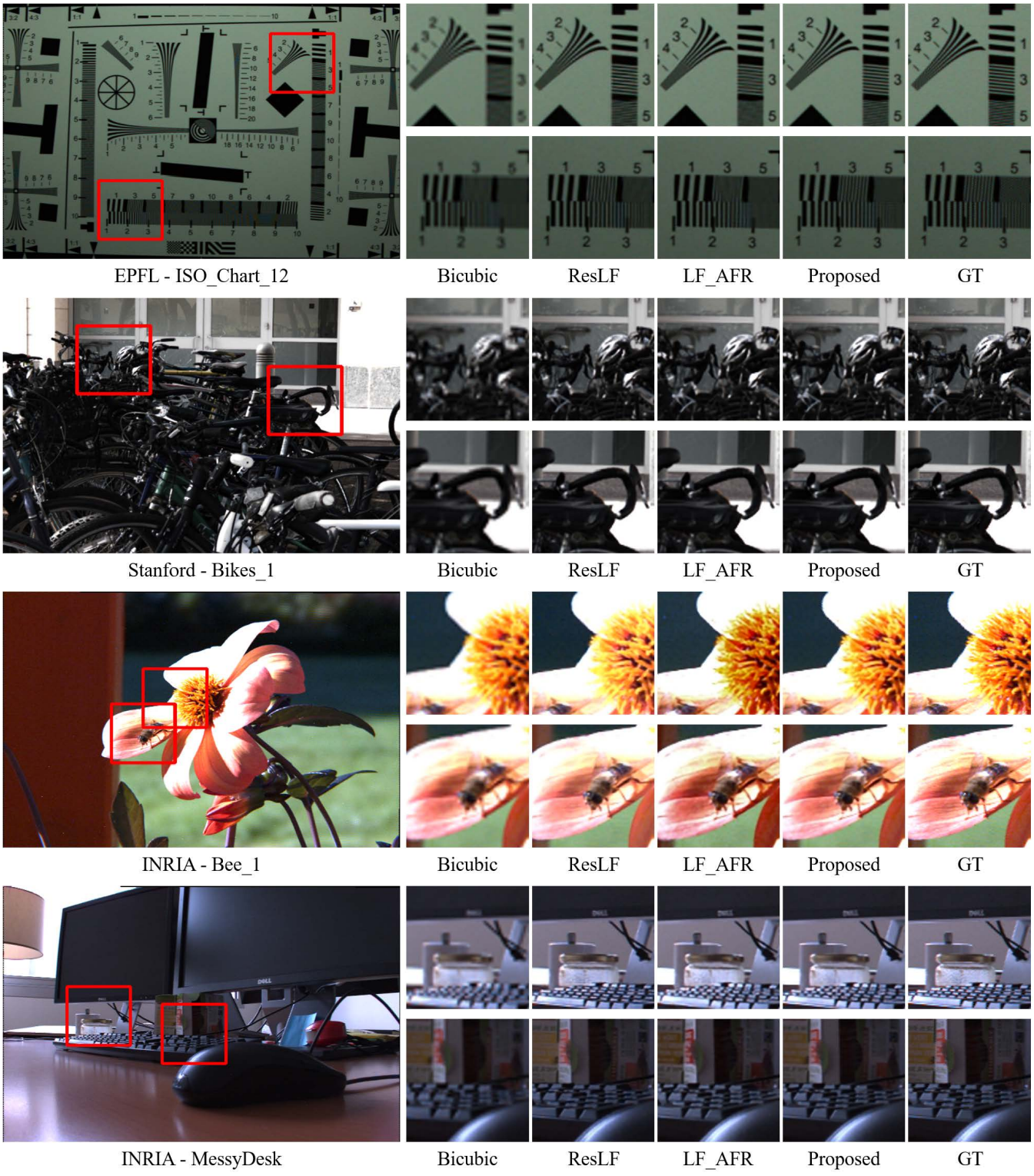

FIGURE 9. Qualitative comparison of the proposed network with Bicubic, ResLF [14], and LFSR_AFR [16] on the real-world datasets.

2) WITHOUT FEATURE WARPING MODULE

We remove the feature warping module from the proposed network to validate the effectiveness of the feature warping module. Without the feature warping module, features of reference images are not aligned to the target image, and thus sub-pixel information between the

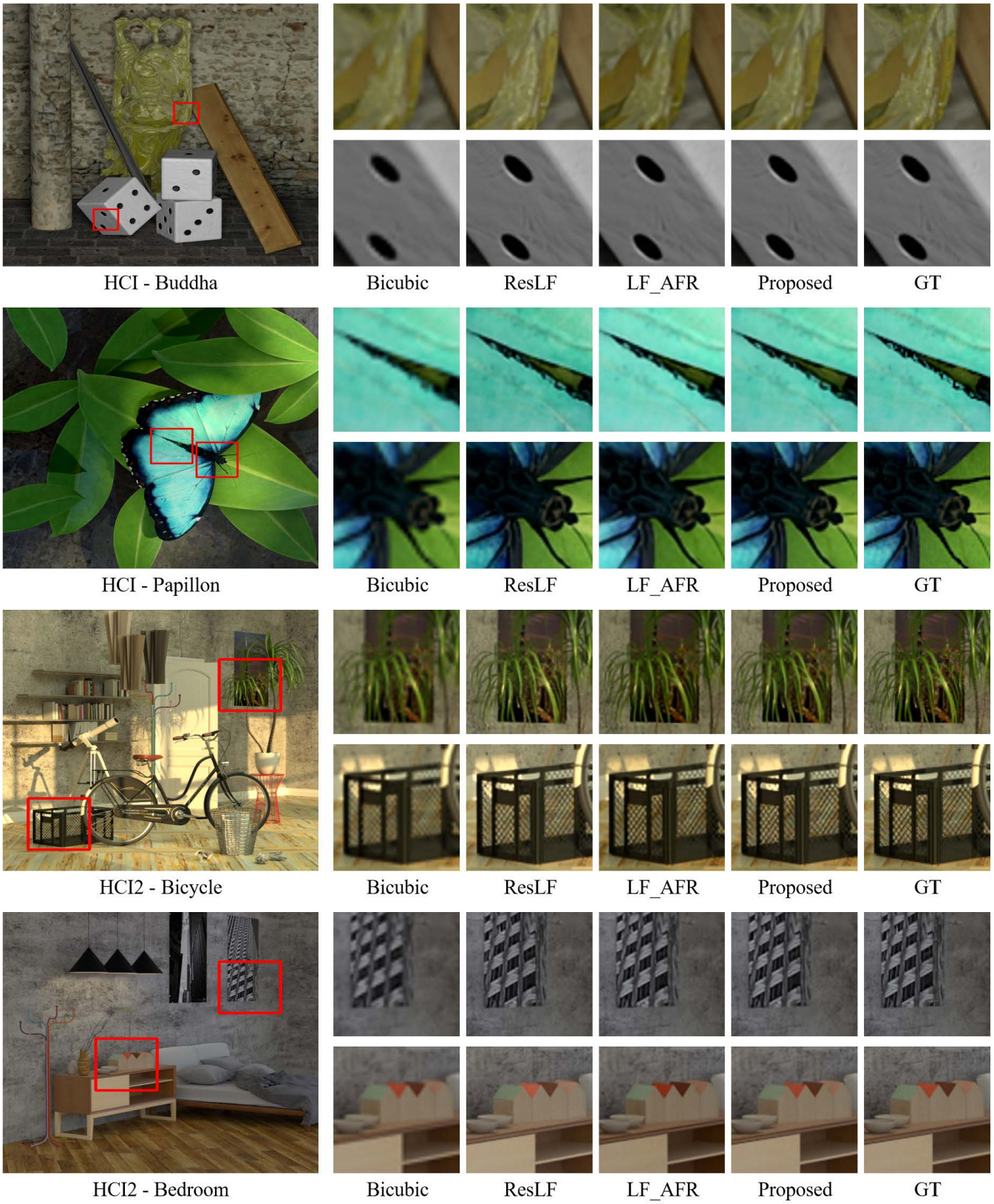

FIGURE 10. Qualitative comparison of the proposed network with Bicubic, ResLF [14], and LFSR_AFR [16] on the synthetic datasets.

TABLE 4. Comparison of the number of network parameters and run time. Here, we super-resolve 187 × 270 × 9 × 9 LF images to 374 × 540 × 9 × 9.

	ResLF	LFSR_AFR	Proposed
#Params./ run time	9.31M/ 2.64s	1.6M/ 4.96s	1.58M/ 3.51s

target and reference images cannot be precisely exploited. To this end, incorrect sub-pixel information degrades the SR performance as in Table 2.

3) WITHOUT FEATURE MIXING

we measure the performance of the proposed network without the feature mixing module.In Table 2, we observe that the performance is degraded severely for all datasets. This indicates that the feature mixing module is designed to combine various features in the target and reference images effectively.

4) WITHOUT MULTI-LEVEL STRUCTURE

we remove all multi-level structures in the proposed network. Specifically, in the feature extractor, only single-level (highest-level) feature is extracted from each view image. Also, in feature warping module, the warping process is performed on the single-level feature for each image. Finally, only single-level mixed feature ($\mathbf{M}^5$) is extracted from the feature mixing module. Then, the upscaling module increases the resolution using the single-level mixed feature. This variation degrades the performance of the proposed network, since receptive fields of various sizes cannot be available.

C. COMPARISON WITH STATE-OF-THE-ARTS

Table 3 and Figure 8 compare the proposed network with the existing LFSR networks (LFNet [33], ResLF [14], and LFSR_AFR [16]), the SISR network (EDSR [42]) the SISR network (EDSR [42]), and the video SR network (SOF-VSR [43]). The PSNR and SSIM scores of the existing algorithms on the HCI, HCI2, EPFL, and Stanford datasets are from LFSR_AFR [16]. For the INRIA datset, we compare SR results using the source codes, provided by the respective authors.

In Table 3, we observe that the proposed network achieves the best performance on all datasets. Especially, the proposed network outperforms the state-of-the-art [16] with significant margins on the real-world LF dataset (EPFL, Stanford, INRIA). Figure 9 and Figure 10 illustrate qualitative LFSR results for real-world and synthetic images, respectively. The proposed network generates visually pleasing results as comapred with the existing networks. Finally, Table 4 shows the number of network parameters and run time of the proposed algorithm and the state-of-the-arts [14], [16]. The proposed network requires the smallest number of parameters. It is worth pointing out that the proposed network surpasses the state-of-the-arts, even though the proposed network uses the minimum network parameters. Also, the proposed network is faster than LFSR_AFR [16].

V. CONCLUSION

In this paper, we proposed the LFSR network based on multi-level structures. The proposed network extracts multi-level features to estimate sub-pixel information from reference images effectively. It then provides multi-level mixed features by combining reference features based on the similarity between the target and reference images. Using multi-level mixed features, the proposed network gradually reconstructs a residual image for SR. Experimental results demonstrated that the proposed network outperforms the state-of-the-art LFSR methods on various LF datasets.

REFERENCES

[1] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, 2005.

[2] N. Liu, W. Zhao, D. Zhang, J. Han, and L. Shao, "Light field saliency detection with dual local graph learning and reciprocative guidance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4712–4721.

[3] T. Wang, Y. Piao, H. Lu, X. Li, and L. Zhang, "Deep learning for light field saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8838–8848.

[4] K. Mishiba, "Fast depth estimation for light field cameras," *IEEE Trans. Image Process.*, vol. 29, pp. 4232–4242, 2020.

[5] A. Chuchvara, A. Barsi, and A. Gotchev, "Fast and accurate depth estimation from sparse light fields," *IEEE Trans. Image Process.*, vol. 29, pp. 2492–2506, 2020.

[6] T. Li, D. P. K. Lun, Y.-H. Chan, and Budianto, "Robust reflection removal based on light field imaging," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1798–1812, Apr. 2019.

[7] M. Le Pendu, X. Jiang, and C. Guillemot, "Light field inpainting propagation via low rank matrix completion," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1981–1993, Apr. 2018.

[8] Y. Wang, T. Wu, J. Yang, L. Wang, W. An, and Y. Guo, "DeOccNet: Learning to see through foreground occlusions in light fields," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 118–127.

[9] S. Wanner, S. Meister, and B. Goldlücke, "Datasets and benchmarks for densely sampled 4D light fields," *Vis., Model. Vis.*, vol. 13, pp. 225–226, Sep. 2013.

[10] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. ACCV*, 2016, pp. 19–34.

[11] *The Stanford Lytro Light Field Archive*. [Online]. Available: http://lightfields.stanford.edu/LF2016.html/

[12] M. Řeřábek and T. Ebrahimi, "New light field image dataset," in *Proc. QoMEX*, 2016.

[13] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–10, Nov. 2016.

[14] S. Zhang, Y. Lin, and H. Sheng, "Residual networks for light field image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11046–11055.

[15] Y. Wang, J. Yang, L. Wang, X. Ying, T. Wu, W. An, and Y. Guo, "Light field image super-resolution using deformable convolution," *IEEE Trans. Image Process.*, vol. 30, pp. 1057–1071, 2020.

[16] K. Ko, Y. J. Koh, S. Chang, and C.-S. Kim, "Light field super-resolution via adaptive feature remixing," *IEEE Trans. Image Process.*, vol. 30, pp. 4114–4128, 2021.

[17] S. Zhang, S. Chang, and Y. Lin, "End-to-end light field spatial super-resolution network using multiple epipolar geometry," *IEEE Trans. Image Process.*, vol. 30, pp. 5956–5968, 2021.

[18] J. Jin, J. Hou, J. Chen, and S. Kwong, "Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2260–2269.

[19] M. S. K. Gul and B. K. Gunturk, "Spatial and angular resolution enhancement of light fields using convolutional neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2146–2159, May 2018.

[20] K. I. Kim and Y. Kwon, "Single-image super-resolution using sparse regression and natural image prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1127–1133, Jun. 2010.
[21] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. ECCV*. Cham, Switzerland: Springer, 2014, pp. 184–199.
[22] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.
[23] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. ECCV*, Sep. 2018, pp. 286–301.
[24] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5690–5699.
[25] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[26] L. Wang, T.-K. Kim, and K.-J. Yoon, "EventSR: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8315–8325.
[27] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3224–3232.
[28] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3360–3369.
[29] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[30] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1954–1963.
[31] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 24–32.
[32] H. Fan, D. Liu, Z. Xiong, and F. Wu, "Two-stage convolutional neural network for light field super-resolution," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1167–1171.
[33] Y. Wang, F. Liu, K. Zhang, G. Hou, Z. Sun, and T. Tan, "LFNet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4274–4286, Sep. 2018.
[34] H. W. F. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, and Y. Y. Chung, "Light field spatial super-resolution using deep efficient spatial-angular separable convolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2319–2330, May 2019.
[35] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 1996, pp. 31–42.
[36] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9308–9316.
[37] B. Li, B. Wang, J. Liu, Z. Qi, and Y. Shi, "S-LWSR: Super lightweight super-resolution network," *IEEE Trans. Image Process.*, vol. 29, pp. 8368–8380, 2020.
[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MIC-CAI*, 2015, pp. 234–241.
[39] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.
[41] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1. Princeton, NJ, USA: Citeseer, 2013, p. 3.
[42] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.
[43] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, "Deep video super-resolution using HR optical flow estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 4323–4336, 2020.

HWA-JONG PARK received the B.S. degree from the Department of Computer Engineering, Korea National University of Transportation, South Korea, in 2020, and the M.S. degree from the Department of Computer Engineering, Chungnam National University, in 2022. He is currently works with Intekplus Company, South Korea. His main research interests include image processing, machine vision, and deep learning.

JUNHO SHIN received the bachelor's degree in computer science and engineering from Chungnam National University, Daejeon, South Korea, in 2021, where he is currently pursuing the master's degree. His current research interests include deep learning, computer vision, especially face recognition and clustering.

HYUNSEOP KIM received the B.S. degree in computer science and engineering from Chungnam National University, Daejeon, South Korea, in 2021, where he is currently pursuing the M.S. degree. His research interest includes computer vision.

YEONG JUN KOH received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2011 and 2018, respectively. He was an Assistant Professor, in March 2019, he joined the Department of Computer Science and Engineering, Chungnam National University. His research interests include computer vision and machine learning, especially in the problems of video object segmentation, and image enhancement.