Light Field Super-Resolution: A Benchmark

Zhen Cheng Zhiwei Xiong* Chang Chen Dong Liu
University of Science and Technology of China

Abstract

Lenslet-based light field imaging generally suffers from a fundamental trade-off between spatial and angular resolutions, which limits its promotion to practical applications. To this end, a substantial amount of efforts have been dedicated to light field super-resolution (SR) in recent years. Despite the demonstrated success, existing light field SR methods are often evaluated based on different degradation assumptions using different datasets, and even contradictory results are reported in literature. In this paper, we conduct the first systematic benchmark evaluation for representative light field SR methods on both synthetic and real-world datasets with various downsampling kernels and scaling factors. We then analyze and discuss the advantages and limitations of each kind of method from different perspectives. Especially, we find that CNN-based single image SR without using any angular information outperforms most light field SR methods even including learning-based ones. This benchmark evaluation, along with the comprehensive analysis and discussion, sheds light on the future researches in light field SR.

1. Introduction

The light field imaging technique enables capture of the light rays not only at different locations but also from different directions [6]. Owing to the redundant spatio-angular information recorded in light field images, many novel applications such as post-capture refocusing [35], stereoscopic display [25], and single-shot depth sensing [36, 37, 38] become possible and popular, especially after the emergence of commercialized portable light field cameras such as Lytro [1] and Raytrix [2]. Despite such advantages of light field imaging, several researches have also pointed out that there is a fundamental trade-off between spatial and angular resolutions [28, 35] that can be obtained. For portable light field cameras, the micro-lens-array placed between the main lens and the sensor plane virtually splits the main lens into sub-apertures, which trades the spatial resolution of the sensor for the angular resolution.

*Correspondence should be addressed to zwxiong@ustc.edu.cn

The spatio-angular resolution trade-off of light field imaging limits its promotion to practical applications. Therefore, light field super-resolution (SR) has drawn more and more attention from researchers and a number of methods have been proposed to take advantage of the redundant information in the 4D light field to solve this problem.[1] Broadly speaking, these light field SR methods can be divided into three categories: projection-based, optimization-based, and learning-based. Relying on the imaging principles of light field cameras, projection-based methods [17, 29, 31, 34] propagate the pixels of each sub-aperture image to the target view by using the abundant sub-pixel information. Optimization-based methods [5, 7, 14, 33, 40, 41, 50] super-resolve the light field under various optimization frameworks with priors analyzed from different mathematical models. Learning-based methods [13, 15, 21, 48, 55, 56] use powerful statistical learning tools especially convolutional neural networks (CNNs) to derive an appropriate mapping from low-resolution (LR) light fields to high-resolution (HR) ones. As a result, the super-resolved images are demonstrated to be useful in light field applications such as disparity estimation [21, 48, 50, 55].

Despite the demonstrated success, existing light field SR methods are usually evaluated with different datasets and the LR images are generated under different degradation assumptions (*i.e.*, downsampling kernels and scaling factors). It makes the comparison among them difficult, and even contradictory results are reported in literature. Therefore, it is of great interest and importance to systematically evaluate existing light field SR methods under a unified setting. For a fair comparison, the test LR light field images together with the ground truth should be the same for every evaluated method, and various degradation assumptions should be considered as well. In addition, due to the notable difference in image quality and scene content between synthetic and real-world light field images, these two kinds of datasets should be both taken into account.

In this paper, we conduct the first systematic benchmark evaluation on several representative light field SR meth-

[1] We mainly discuss the spatial resolution enhancement in this paper, although the angular resolution enhancement (*i.e.*, view synthesis) has also been frequently investigated [24, 47, 52, 53].

ods. We select two datasets that are commonly used in light field researches for the evaluation. One is the HCI synthetic dataset [51] and the other is the EPFL real-world dataset [39]. We then examine the performance of all considered light field SR methods under degradation assumptions with various downsampling kernels and scaling factors. The HR light field images generated by each method are evaluated using four image quality metrics in terms of both reconstruction accuracy and perceptual quality. According to the evaluation results, we then analyze and discuss the advantages and limitations of each kind of method from different perspectives.

Especially, besides the representative light field SR methods, we also adopt a CNN-based single image SR method without using any angular information in the light field for an additional comparison, which outperforms most light field SR methods even including learning-based ones. It is not so surprising since the single image SR method relies on a powerful 2D natural image prior learned from a large external dataset, while the light field SR methods either exploit the inter-view redundancy within the LR input only or learn from limited external data. However, it reveals that there is still a large room of improvement for light field SR. Based on this observation as well as the above analysis and discussion, we point out the key challenges for light field SR and further propose several promising directions to address them.

Contributions of this paper are highlighted as follows:

(1) The first benchmark evaluation for light field SR.

(2) Interesting and informative observations on the advantages and limitations of representative methods in terms of different datasets, degradations, and metrics.

(3) Analysis and discussion for promising directions in future light field SR researches.

2. Related Work

In general, light field SR aims to enhance the spatial resolution of each sub-aperture image from an LR light field by exploiting the redundant information across the angular dimensions. Existing methods can be broadly divided into three categories: projection-based, optimization-based, and learning-based. Note that methods using additional hardware [9, 59, 60] are not included.

Projection-based methods rely on the imaging principles of light field cameras. As first introduced by Lim *et al*. [31], the 2D sub-aperture images contain sub-pixel shift in the spatial dimensions, which can be used for spatial resolution enhancement after projecting them onto convex sets. Nava *et al*. [34] exploited the refocusing principle and projected pixels from other views to the central view to get an all-in-focus image of the present scene. Georgiev *et al*. [17] also established sub-pixel correspondences with the projection scheme designed for the focused plenotic cameras. Liang *et al*. [29] proved that typical lenslet light field cameras preserve frequency components above the spatial Nyquist rate and projected the light field samples to the target view with the guidance of scene depth to make use of the redundant inter-view information.

Optimization-based methods utilize various optimization frameworks to super-resolve light field images, relying on different mathematical or geometric modeling of the 4D light field structure. Bishop *et al*. [7] explicitly introduced Lambertian reflectance and texture preserving priors in the light field imaging model and reconstructed the HR light field images with a variational Bayesian framework. Mitra and Veeraraghavan [33] assumed that the disparity is constant within each 4D light field patch and estimated the HR light field patches using a linear minimum mean square error estimator with a disparity-dependent Gaussian mixture model. Wanner and Goldluecke [50] applied a variational framework to conduct both spatial and angular SR using the disparity maps estimated from the epipolar images (EPIs) with a structure tensor method. Recently, Rossi *et al*. [41] super-resolved the light fields by coupling the multi-frame approach with a graph-based regularizer that enforces the light field structure and avoids explicit disparity estimation. Their subsequent work [40] prevented the low-pass tendency of the quadratic regularizer by replacing it with a nonsmooth square root regularizer. Inspired by LFBM5D for light field denoising [4], Alain and Smolic [5] proposed an algorithm that iteratively alternates between LFBM5D filtering and back-projection for light field SR.

Learning-based methods emerge recently especially due to the prosperity of deep learning. Farrugia *et al*. [15] showed that the light field patch volume resides in a low-dimensional subspace and learned a linear mapping between the LR and HR subspaces with ridge regression. Deep learning for light field SR was first introduced by Yoon *et al*. in [55], where they stacked 4-tuples of sub-aperture images and fed them into the SRCNN [11] architecture with multiple channels. Fan *et al*. [13] developed a two-stage CNN framework, where different sub-aperture images are aligned by patch matching in the first stage and a multi-patch fusion CNN is used in the second stage. A shallow CNN was proposed by Gul *et al*. [21] to super-resolve light fields directly from the raw data captured by plenoptic cameras without decoding to sub-aperture images. Considering a light field as a sequence of 2D images, Wang *et al*. [48] modeled the spatial correlation between adjacent views with a bidirectional recurrent CNN and accumulated contextual information from multiple scales with a specially designed fusion layer. With a combined CNN architecture, Yuan *et al*. [56] performed light field SR with the EDSR [30] network followed by an EPI enhancement network.

3. Benchmark Settings

3.1. Datasets

We select two datasets that are widely used in light field researches for the benchmark evaluation. The HCI dataset [51] originally proposed for light field disparity estimation [22, 45] contains a number of scenes synthesized by graphic software. To facilitate a fair comparison especially for learning-based methods, we select 10 scenes with a uniform angular resolution of 9×9 and the spatial resolution ranging from 768×768 to 1024×720. The EPFL dataset [39] originally proposed for light field image compression [12] contains 12 real-world scenes captured by the Lytro Illum camera. The resolution of these 4D light field images is $625 \times 434 \times 15 \times 15$. Unlike the synthetic images, the real-world images suffer from the vignetting effect even after calibration with the built-in camera firmware [10]. Therefore, we only use the central 9×9 views from the original light field and conduct a further rectification by matching the average intensity of each sub-aperture image to that of the central view. Note that even after the post-processing, the real-world light field may still have view-dependent camera degradations such as noise, as will be seen in the experiments. More details of the datasets are provided in the supplementary document.

3.2. Degradations

There are several ways to simulate the degradation from HR light field images to LR ones. With different degradation assumptions, the resulting LR light field images as well as their interpolated ones may be drastically different. This poses the main difficulty for directly comparing existing light field SR methods since their inputs may be drastically different even the same groud truth is used. Without loss of generality, we simulate four degradation models with two different downsampling kernels (Bicubic and Gaussian) and two different scale factors (2 and 3). For Bicubic downsampling, we use the MATLAB function *imresize*. For Gaussian downsampling, we blur each ground truth sub-aperture image using a 3×3 Gaussian kernel with the standard deviation of 2. The LR light field is obtained by averaging the neighboring four pixels for the scale factor of 2 and directly sampling the central pixel for the scale factor of 3 in each blurred sub-aperture images. In this way, the LR light field can be well aligned with the HR ground truth after interpolation.

3.3. Methods

We select four representative light field SR methods from the three categories as mentioned above for the evaluation based on the following considerations: 1) recently proposed state-of-the-art (within the last three years), 2) publicly available or easily implementable codes, and 3) generalizability to different downsampling kernels and scaling factors. Specifically, we adopt the projection-based method (PRO) [29] using estimated scene depth by [45], the optimization-based method using graphs (GB) [41], the learning-based method with ridge regression (RR) [15], and the first CNN-based method (LFCNN) [55]. For an additional comparison, we also adopt a representative CNN-based single image SR method (VDSR) [26] without using any angular information in the light field but trained from a large external 2D image dataset. These methods are either implemented using the author provided codes [15, 41] or that developed by ourselves [26, 29, 55]. In both cases, they are validated with the results in the original paper. It is worth mentioning that, despite our best efforts, each selected method may not give the top performance in the category which it belongs to, yet the overall picture drawn from the experiments should still hold. Also, interested researchers can easily add their own methods that may give better results into this benchmark evaluation.

Method	Language	Category	Time (s)
BIC	MATLAB	Single image	0.002
PRO [29]	MATLAB&C++	Projection	113.0
GB [41]	MATLAB	Optimization	286.9
RR [15]	MATLAB	Learning	24.02
LFCNN [55]	MATLAB&C++	Learning	0.036
VDSR [26]	MATLAB&C++	Single image	0.138

Table 1. Evaluated methods and average execution time for super-resolving one sub-aperture image from an input light field with a $256 \times 256 \times 9 \times 9$ resolution. The execution time is measured under Gaussian downsampling with the scale factor of 3 on a machine with a 3.2GHz CPU (for BIC, PRO, GB, and RR) and a GTX 1080Ti GPU (for LFCNN and VDSR). More details of the evaluated methods are provided in the supplementary document.

Table 1 lists the implementation language of these methods along with their categories and average execution time. Bicubic interpolation (BIC) is included as the baseline. For PRO [29] and GB [41] that involve several tunable parameters, we select the setting that gives the best results. For RR [15], we use the PCA basis and transformation matrices learned from an additional dataset without overlap of our test dataset, which are provided by the authors. For LFCNN [55] which need to be trained on part of the dataset, we use the K-fold cross validation strategy to get the SR results on the whole dataset. Specifically, we split each dataset to test groups with 2 (for HCI dataset) or 3 (for EPFL dataset) scenes in each group and use the scenes outside each group to train the network model. Note that we upgrade the shallow SRCNN structure originally used in LFCNN to the deep VDSR structure, which promotes its performance for a fair comparison with single image VDSR. For single image VDSR, we train the network using the same training set as in [26] under different degradations.

3.4. Metrics

We use the PSNR and SSIM [49] metrics to evaluate the reconstruction accuracy. Besides, considering the tradeoff between reconstruction accuracy and perceptual quality as revealed in [8], we also use the VGG metric [58] and Ma's metric [32] to evaluate the perceptual quality besides direct visual comparison. It is worth mentioning that, given an LR light field, not all light field SR methods output a complete 4D HR light field at once. Specifically, LFCNN [55] uses a 4-tuple of sub-aperture images as the input and the output is still a 4-tuple of HR sub-aperture images. We can repeat this process to obtain the whole HR light field. For PRO [29], however, it only generates the HR central view. Therefore, we conduct the evaluation on the super-resolved central view image for all methods and on all sub-aperture images except PRO. For the latter, we compare the average results and their standard deviation over all views.

4. Results and Analysis

4.1. Reconstruction accuracy evaluation

Comparison to baseline. Fig. 1 plots the average PSNR values of the super-resolved central view images for six selected methods over two datasets and under four degradation models, which gives an overall picture of this benchmark evaluation. At the first glance, all light field SR methods outperform the baseline BIC in all cases, which demonstrates the effectiveness of exploiting the inter-view information. We cannot take this for granted. Actually, considering the complicated scene content especially occlusion present in the sub-aperture images, it is possible that inferior results to BIC could be generated if the inter-view information is not properly used. Therefore, the advantages of these selected methods are validated comprehensively. The results in terms of the SSIM metric [49] are provided in the supplementary document, from which we have similar observations.

Synthetic dataset VS. real-world dataset. We further divide the four light field SR methods into two groups, non-learning-based including PRO and GB, and learning-based including RR and LFCNN. We observe that, except for Gaussian downsampling with the scale factor of 2 (as will be explained below), non-learning-based methods give competitive or even better results to learning-based ones on the HCI synthetic dataset, while on the EPFL real-world dataset, learning-based methods have a clear advantage. The underlying reason is that, compared with the real-world light fields, the synthetic light fields generally has much cleaner and simpler scene content, which facilities the projection-based and optimization-based methods that rely on system or mathematic modeling of light field imaging. In contrast, learning-based methods are more robust even for noisy or cluttered scene content in real world.

Bicubic downsampling VS. Gaussian downsampling. In terms of the degradation model, a notable thing is that LFCNN significantly outperforms the other ones for Gaussian downsampling with the scale factor of 2. Recall that this degradation model is operated as first blurring the sub-aperture image with a Gaussian kernel and then averaging the four neighboring pixels for downsampling. Compared with other degradation models, this one actually conducts twice of low-pass filtering and thus results in more heavily degraded LR images, which can be verified by the baseline BIC results. This degradation is thus more challenging for non-learning-based methods PRO and GB and even the conventional learning-based method RR, while LFCNN stands out owing to the power of deep learning.

Light field SR VS. single image SR. Besides the light field SR methods, we also evaluate the performance of single image VDSR without using the inter-view information. As can be seen from Fig. 1, single image VDSR almost always gives the best performance among its competitors including LFCNN. This seemingly surprising result is actually reasonable, since single image VDSR relies on a powerful 2D natural image prior learned from a large external dataset, while its competitors either exploit the inter-view information within the LR input only or learn from limited external data (*e.g.*, 8 or 9 scenes for LFCNN). In this sense, this is not a really fair comparison. However, it reveals that these is still a large room of improvement for light field SR, as we will discuss in the following section.

4.2. Perceptual quality evaluation

Perceptual metric. According to [8], there exists a tradeoff between reconstruction accuracy and perceptual quality for image restoration problems. To evaluate the perceptual quality of different methods in a quantitative manner, we adopt the VGG metric [58] that represents the pixel-wise distance in the feature space of a VGG19 network [42]. Fig. 2 plots the average VGG values of the super-resolved central view images for six selected methods over two datasets and under four degradation models. As can be seen, the basic trend is similar to that of the PSNR metric. All light field methods outperform the baseline BIC, and single image VDSR gives best results. However, there are still several notable differences. First, PRO gives promising performance in terms of the VGG metric. For example, under Gaussian downsampling with the scale factor of 3, PRO significantly outperforms LFCNN on the synthetic dataset and even outperforms LFCNN on the real-world dataset where it has a much lower PSNR. Second, RR seems to be not favorable by the VGG metric, even on the real-world dataset where it always outperforms PRO and GB in terms of PSNR. These observations confirm to the perceptual-distortion tradeoff [8]. The results of Ma's metric [32] are provided in the supplementary document.

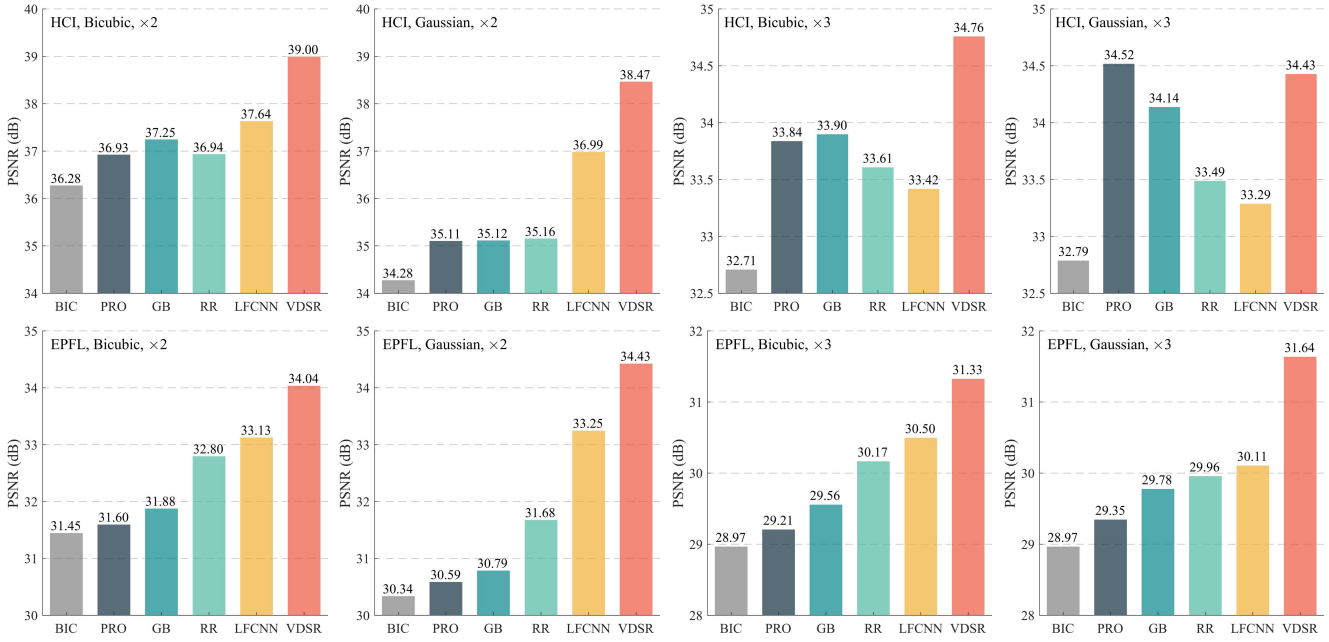

Figure 1. The average PSNR values (the higher, the better) of the super-resolved central view images for six selected methods over two datasets and under four degradation models. The results in terms of the SSIM metric [49] are provided in the supplementary document.

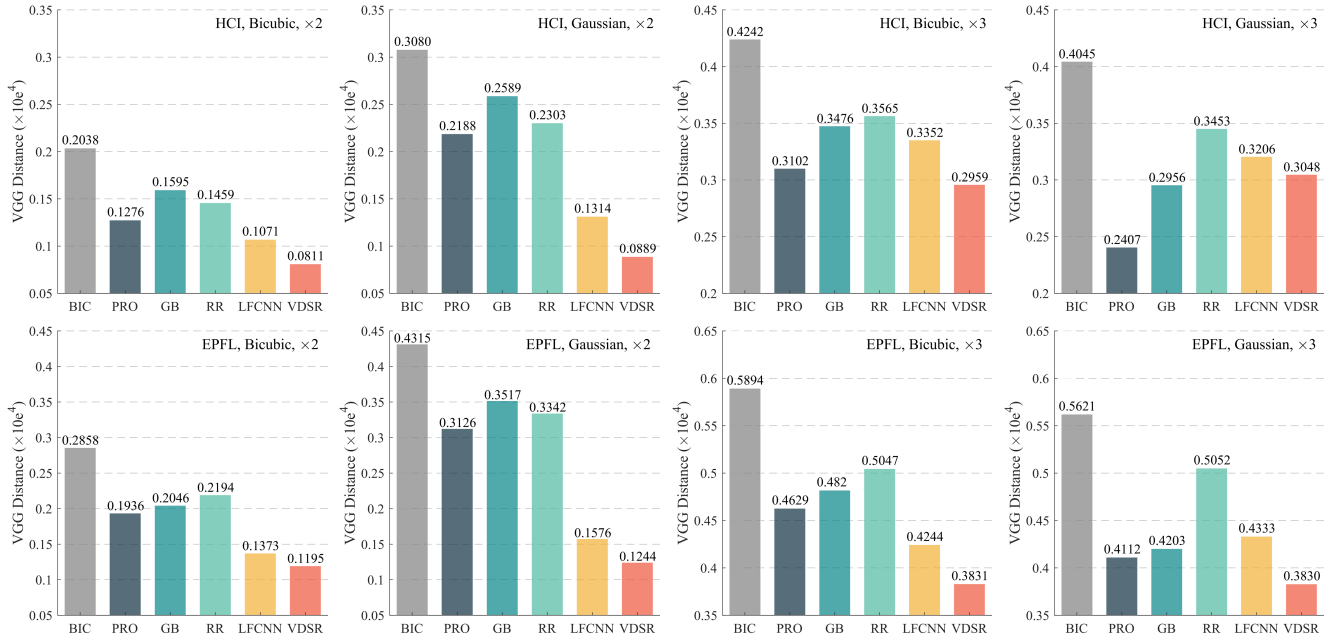

Figure 2. The average VGG values (the lower, the better) of the super-resolved central view images for six selected methods over two datasets and under four degradation models. The results of Ma's perceptual metric [32] are provided in the supplementary document.

Visual results. Fig. 3 shows some super-resolved central view images under Gaussian downsampling with the scale factor of 3 for a qualitative comparison among different methods. We can see that, while all methods add more visual details over the baseline BIC, their behaviors are quite different. Specifically, PRO gives quite impressive results in regions with fine textures and continuous depth (marked in red rectangles) but is not so effective in regions with distinct edges and occlusion (marked in green rectangles). In contrast, LFCNN produces sharp edges but often introduces unrealistic artifacts in texture regions. The visual results from GB are somewhat between PRO and LFCNN

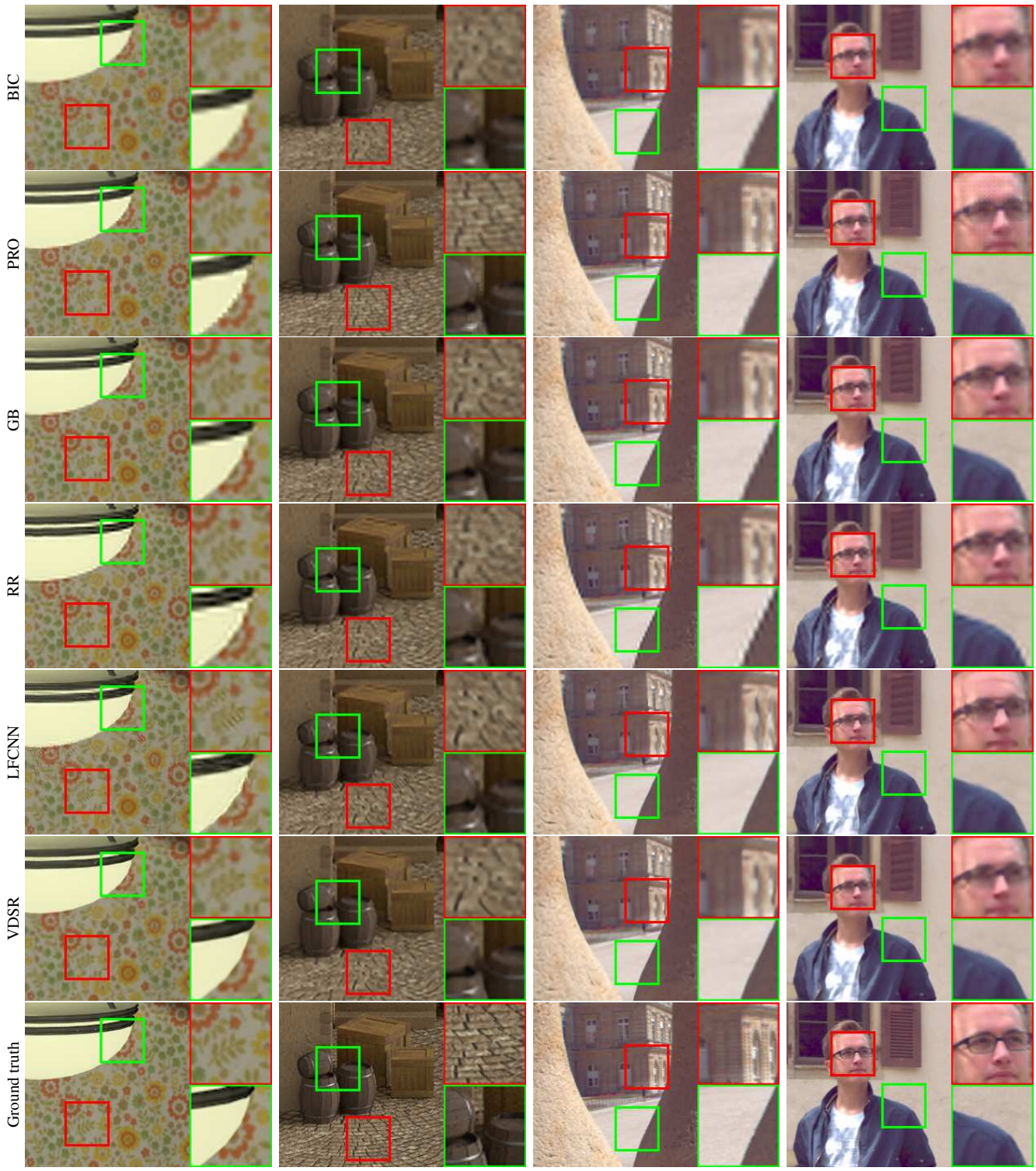

Figure 3. Visual comparisons of super-resolved central view images through different methods under Gaussian downsampling with the scale factor of 3. The first two scenes are from the HCI synthetic dataset and the last two scenes are from the EPFL real-world dataset.

while those from RR are not so encouraging. This is in accordance with the VGG metric, on which PRO and LFCNN are the top two performers in most cases. Besides, single image VDSR without using inter-view information gener-

Method	HCI				EPFL			
	Bicubic×2	Gaussian×2	Bicubic×3	Gaussian×3	Bicubic×2	Gaussian×2	Bicubic×3	Gaussian×3
BIC	36.26 ± 0.07	34.30 ± 0.06	32.75 ± 0.06	32.84 ± 0.07	31.67 ± 0.75	30.56 ± 0.66	29.22 ± 0.71	29.22 ± 0.73
	20.32 ± 0.51	30.84 ± 0.66	42.35 ± 0.93	40.45 ± 0.95	27.42 ± 1.35	41.80 ± 1.73	57.25 ± 2.21	54.53 ± 2.22
GB	37.18 ± 0.21	35.14 ± 0.11	33.89 ± 0.08	34.14 ± 0.11	32.13 ± 0.82	31.04 ± 0.74	29.78 ± 0.81	29.98 ± 0.84
[41]	15.12 ± 1.52	25.15 ± 1.19	34.12 ± 1.04	28.90 ± 1.01	19.08 ± 1.80	33.51 ± 2.14	46.45 ± 2.18	40.78 ± 1.85
RR	36.89 ± 0.15	35.13 ± 0.11	33.62 ± 0.08	33.52 ± 0.09	32.75 ± 0.43	31.68 ± 0.44	30.21 ± 0.54	30.00 ± 0.59
[15]	14.57 ± 0.62	23.55 ± 0.77	35.77 ± 0.93	34.73 ± 1.02	20.91 ± 1.30	32.91 ± 1.59	48.26 ± 2.11	48.26 ± 2.17
LFCNN	37.60 ± 0.12	37.03 ± 0.12	33.46 ± 0.10	33.38 ± 0.13	33.33 ± 0.51	33.53 ± 0.46	30.72 ± 0.55	30.28 ± 0.67
[55]	10.66 ± 0.37	13.00 ± 0.52	33.70 ± 0.98	32.39 ± 1.05	12.78 ± 0.94	14.61 ± 1.07	41.09 ± 2.03	42.06 ± 2.32

Table 2. Mean and standard deviation values of PSNR (dB) and VGG ($\times 100$, displayed in gray) on all sub-aperture images.

Train / Test	Bicubic ×2	Gaussian ×2	Bicubic ×3	Gaussian ×3
EPFL / EPFL	33.13	33.25	30.50	30.11
HCI / EPFL	32.18	31.97	28.96	28.89

Table 3. Domain shift investigation. Average PSNR (dB) results of super-resolving central view images from the EPFL dataset with LFCNN trained on different datasets.

ates even sharper edges than LFCNN, yet PRO still has an advantage in texture regions. The reason is that fine textures are more difficult to learn from external examples, but relatively easy to be enhanced by the internal correlation across the angular dimensions of the light field itself. This suggests a potential way for combining non-learning-based methods and learning-based ones.

4.3. Inter-view consistency

The above comparisons are conducted on the super-resolved central view images. In Table 2 we list the PSNR and VGG results on all sub-aperture images in terms of the mean and standard deviation values, for light field SR methods except PRO which only generates the HR central view. The mean PSNR and VGG values on all sub-aperture images are similar to those on the central view, which suggests the universal effectiveness of selected methods. On the other hand, the standard deviation results are more informative, which indicates the inter-view consistency of each method. As can be seen, BIC has a relatively small deviation on the synthetic dataset but a much larger one on the real-world dataset. This can be explained by the fact that the synthetic light fields have no intensity variation among views but the real-world ones are affected by vignetting as well as other view-dependent camera degradations such as noise. Therefore, the individual operation like BIC will inherit the inter-view variation on real-world data, while this variation could be alleviated by light field SR methods that operate on all sub-aperture images simultaneously. Among the three light field SR methods, RR exhibits the smallest deviation in terms of the PSNR metric and LFCNN exhibits the smallest deviation in terms of the VGG metric.

4.4. Generalizability of LFCNN

Recall that for LFCNN, we use the K-fold cross validation strategy to get the SR results on the whole dataset, which requires to train the model on part of the dataset. Here we conduct another experiment for LFCNN to evaluate its generalization capability from one dataset to another. Specifically, we select one LFCNN network trained on the synthetic dataset and then apply it to the real-world dataset. The PSNR results on the central view images for different degradation models are listed in Table 3. We can see that, due to the domain shift issue, the performance of LFCNN deteriorates by an average of 1.25dB in terms of PSNR. It thus reveals the shortcoming of CNN-based methods for light field SR, despite its superior performance over other methods when trained on part of the dataset. On the other hand, however, it can be expected that LFCNN would benefit from increased training data in the same domain.

4.5. Computational complexity

The computational complexity of the evaluated methods are included in Table 1. While these methods are implemented based on different languages and hardware, it is obvious that non-learning-based method consumes more time than learning-based ones. Owing to the parallel computation, LFCNN only requires 0.036s for super-resolving one sub-aperture image from an input light field with a $256 \times 256 \times 9 \times 9$ resolution in the test phase, although it takes about 7 hours for training the network. Note that LFCNN is averagely faster than VDSR since it processes a 4-tuple of sub-aperture images at once.

5. Light Field SR: Next Step

5.1. Combining natural image priors

Natural image priors such as edge structure and patch recurrence are widely used for single image SR in an early stage [18, 57]. Recent CNN-based methods achieve excellent results with deep learning from a large external image dataset, and the performance can be further improved if trained on 2D images with very high spatial resolution [43]. Although these powerful network structures can be read-

ily extended to light field SR [55], a sufficiently large light field dataset containing diverse content is not easy to collect compared with 2D natural images. On the other hand, the real-world data collected with portable light field cameras often suffers from limited spatial resolution, which also restricts the capability of deep learning. Consequently, as demonstrated above, single image VDSR easily outperforms LFCNN, since it uses much more training data.

While it is definitely necessary to pay efforts for collecting even larger high quality light field datasets than the existing ones, an alternative way for boosting the performance of light field SR could be directly taking advantage of natural image priors. For example, in a most simple manner, the single image SR results can be used as initializations for light field SR [13, 56]. In an advanced manner, single image SR methods that exploit the intra-view information can be combined with light field SR methods that exploit the inter-view information, where they may find complementary strengths [54]. In a word, the powerful natural image priors can be better leveraged.

5.2. Taking full use of 4D structure

Owing to its high dimensional property, light fields enable novel applications beyond conventional 2D images. For light field SR, a main problem is how to take full use of the 4D structure. Take deep-learning-based methods for example, it is essential to utilize network structures that are specially designed for high dimensional data. For instance, 3D CNN was first used for human action recognition from videos [23] and has been proven efficient for integrating spatial and temporal information. In addition, there are some CNN architectures designed for 4D light fields such as the pseudo 4D CNN which is used for view synthesis [47] and the 4D filter mimicked by interleaved spatial and angular filters which is used for material recognition [46]. Both of these two 4D CNN structures outperform traditional 2D CNNs for their specific applications with light fields. It thus reveals a potential way to develop light field SR methods with advanced CNN structures taking full use of the 4D correlations across both spatial and angular dimensions.

5.3. Alleviating domain shift

Since light field imaging was first introduced by [6], the acquisition systems have been developed in different principles such as computer graphics tools [51], lenslet cameras [35], and camera arrays [3]. These systems are quite different from each other. Even within the same category, *e.g.*, lenslet cameras, different configurations of the microlens-array may result in different camera models. This difference, regarded as domain shift, is considerably larger compared with conventional 2D images captured with different devices. As demonstrated by the experiments on LFCNN, although deep learning opens the door for a better modeling of light field SR, the model learned from a certain light field dataset cannot be easily applied to another dataset obtained with different camera configurations. This is another key issue that makes light field SR more challenging than single image SR.

We suggest two possible solutions for addressing this issue. From the perspective of modeling, non-learning-based methods are not so sensitive to the domain shift effect, which indicates that we can incorporate light field modeling used in these methods into deep learning architectures. Take the optimization-based method GB [41] as an example, light fields can be represented as a graph in which one single node represents several rays coming from the same scene point. Meanwhile, graph convolution networks (GCNs) have shown excellent capability in characterizing the relationship between adjacency nodes in many tasks such as classification [27] and 3D shape analysis [44]. Therefore, modeling light fields as graphs and exploiting GCNs to learn the 4D correlations may be a useful solution to address the domain shift issue for light field SR.

On the other hand, from the perspective of domain transfer, domain adaptation techniques are specified for learning tasks in which data at training and testing phases come from similar but different distributions [16]. These techniques are also applicable to alleviate the domain shift effect in light field SR. For example, we can extract the shared features between different light field datasets with adversarial training at a certain layer of the CNNs, which has demonstrated promising domain adaptation performance for object recognition [20] and classification [19].

6. Conclusion

In this paper, the first benchmark evaluation is conducted for light field SR. We systematically evaluate the performance of representative light field SR methods on two sets of light field images for synthetic and real-world scenes under various degradation assumptions. Comprehensive experimental results and further analysis reveal the advantages and limitations of these methods. Based on the benchmark evaluation and corresponding analysis, we suggest several promising directions for the development of more effective methods in the future. We hope this benchmark along with the discussion will not only provide a clear picture for the current status of light field SR but also inspire novel ideas in this important field.

Acknowledgements

We acknowledge funding from National Key R&D Program of China under Grant 2018YFC0307905, Natural Science Foundation of China (NSFC) under Grant 61671419, and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDB06040900.

References

[1] *https://www.lytro.com/*. 1

[2] *https://www.raytrix.de/*. 1

[3] *http://lightfield.stanford.edu/lfs.html/*. 8

[4] M. Alain and A. Smolic. Light field denoising by sparse 5d transform domain collaborative filtering. In *MMSP*, 2017. 2

[5] M. Alain and A. Smolic. Light field super-resolution via lfbm5d sparse coding. In *ICIP*, 2018. 1, 2

[6] J. Bergen and E. Adelson. The plenoptic function and the elements of early vision. *Computational Models of Visual Processing*, 1991. 1, 8

[7] T. E. Bishop and P. Favaro. The light field camera: Extended depth of field, aliasing, and superresolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):972–986, 2012. 1, 2

[8] Y. Blau and T. Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018. 4

[9] V. Boominathan, K. Mitra, and A. Veeraraghavan. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In *ICCP*, 2014. 2

[10] D. G. Dansereau, O. Pizarro, and S. B. Williams. Decoding, calibration and rectification for lenselet-based plenoptic cameras. In *CVPR*, 2013. 3

[11] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 2

[12] T. Ebrahimi, S. Foessel, F. Pereira, and P. Schelkens. Jpeg pleno: Toward an efficient representation of visual reality. *IEEE Multimedia*, 23(4):14–20, 2016. 3

[13] H. Fan, D. Liu, Z. Xiong, and F. Wu. Two-stage convolutional neural network for light field super-resolution. In *ICIP*, 2017. 1, 2, 8

[14] S. Farag and V. Velisavljevic. A novel disparity-assisted block matching-based approach for super-resolution of light field images. In *3DTV-CON*, 2018. 1

[15] R. A. Farrugia, C. Galea, and C. Guillemot. Super resolution of light field images using linear subspace projection of patch-volumes. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):1058–1071, 2017. 1, 2, 3, 7

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 8

[17] T. Georgiev, G. Chunev, and A. Lumsdaine. Superresolution with the focused plenoptic camera. In *SPIE Proc.*, 2011. 1, 2

[18] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *ICCV*, 2009. 7

[19] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011. 8

[20] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011. 8

[21] M. S. K. Gul and B. K. Gunturk. Spatial and angular resolution enhancement of light fields using convolutional neural networks. *IEEE Transactions on Image Processing*, 27(5):2146–2159, 2018. 1, 2

[22] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. So Kweon. Accurate depth map estimation from a lenslet light field camera. In *CVPR*, 2015. 3

[23] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013. 8

[24] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics*, 35(6), 2016. 1

[25] A. Katayama. A view point dependent streoscopic display using interpolation of multi-viewpoint images. In *SPIE Proc.*, 1995. 1

[26] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 3

[27] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2016. 8

[28] A. Levin, W. T. Freeman, and F. Durand. Understanding camera trade-offs through a bayesian analysis of light field projections. In *ECCV*, 2008. 1

[29] C.-K. Liang and R. Ramamoorthi. A light transport framework for lenslet light field cameras. *ACM Transactions on Graphics*, 34(2):16:1–16:19, 2015. 1, 2, 3, 4

[30] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 2

[31] J. Lim, H. Ok, B. Park, J. Kang, and S. Lee. Improving the spatail resolution based on 4d light field data. In *ICIP*, 2009. 1, 2

[32] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 4, 5

[33] K. Mitra and A. Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. In *CVPRW*, 2012. 1, 2

[34] F. P. Nava and J. P. Luke. Simultaneous estimation of super-resolved depth and all-in-focus images from a plenoptic camera. In *The True Vision-Capture, Transmission and Display of 3D Video*, 2009. 1, 2

[35] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report, Stanford University*, 2(11):1–11, 2005. 1, 8

[36] J. Peng, Z. Xiong, D. Liu, and X. Chen. Unsupervised depth estimation from light field using a convolutional neural network. In *International Conference on 3D Vision*, 2018. 1

[37] J. Peng, Z. Xiong, Y. Zhang, D. Liu, and F. Wu. Lf-fusion: Dense and accurate 3d reconstruction from light field images. In *VCIP*, 2017. 1

[38] C. Perwass and L. Wietzke. Single lens 3d-camera with extended depth-of-field. In *Human Vision and Electronic Imaging*, 2012. 1

[39] M. Rerabek and T. Ebrahimi. New light field image dataset. In *International Conference on Quality of Multimedia Experience (QoMEX)*, 2016. 2, 3

[40] M. Rossi, M. El Gheche, and P. Frossard. A nonsmooth graph-based approach to light field super-resolution. In *ICIP*, 2018. 1, 2

[41] M. Rossi and P. Frossard. Graph-based light field super-resolution. In *MMSP*, 2017. 1, 2, 3, 7, 8

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[43] R. Timofte, E. Agustsson, L. V. Gool, and et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 7

[44] N. Verma, E. Boyer, and J. Verbeek. Feastnet: Feature-steered graph convolutions for 3d shape analysis. In *CVPR*, 2018. 8

[45] T.-C. Wang, A. A. Efros, and R. Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *ICCV*, 2015. 3

[46] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi. A 4d light-field dataset and cnn architectures for material recognition. In *ECCV*, 2016. 8

[47] Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, and T. Tan. End-to-end view synthesis for light field imaging with pseudo 4dcnn. In *ECCV*, 2018. 1, 8

[48] 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 Transactions on Image Processing*, 27(9):4274–4286, 2018. 1, 2

[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4, 5

[50] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):606–619, 2014. 1, 2

[51] S. Wanner, S. Meister, and B. Goldlücke. Datasets and benchmarks for densely sampled 4d light fields. In *International Symposium on Vision Modeling and Visualization*, 2013. 2, 3, 8

[52] H. Wing Fung Yeung, J. Hou, J. Chen, Y. Ying Chung, and X. Chen. Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In *ECCV*, 2018. 1

[53] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu. Light field reconstruction using deep convolutional network on epi. In *CVPR*, 2017. 1

[54] Z. Xiong, Z. Cheng, J. Peng, H. Fan, D. Liu, and F. Wu. Light field super-resolution using internal and external similarities. In *ICIP*, 2017. 8

[55] 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 *ICCVW*, 2015. 1, 2, 3, 4, 7, 8

[56] Y. Yuan, Z. Cao, and L. Su. Light-field image superresolution using a combined deep cnn based on epi. *IEEE Signal Processing Letters*, 25(9):1359–1363, 2018. 1, 2, 8

[57] L. Zhang and X. Wu. An edge-guided image interpolation algorithm via directional filtering and data fusion. *IEEE Transactions on Image Processing*, 15(8):2226–2238, 2006. 7

[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4

[59] M. Zhao, G. Wu, Y. Li, X. Hao, L. Fang, and Y. Liu. Cross-scale reference-based light field super-resolution. *IEEE Transactions on Computational Imaging*, 4(3):406–418, 2018. 2

[60] H. Zheng, M. Guo, H. Wang, Y. Liu, and L. Fang. Combining exemplar-based approach and learning-based approach for light field super-resolution using a hybrid imaging system. In *ICCVW*, 2017. 2