RESEARCH

Depth-Assisted Super-Resolution of Light Field in Layered Object Space

Yongwei Li and Mårten Sjöström*

*Correspondence:
marten.sjostrom@miun.se
Department of Information Systems and Technology, Mid Sweden University, Sundsvall, Sweden
Full list of author information is available at the end of the article

Abstract

The captured light field may fail to reconstruct fine details of the scene due to under-sampling problem of light field acquisition devices. Therefore, super-resolution is required to restore high-frequency information from the light field and to improve the quality of the rendered views. Conventional super-resolution algorithms are not ideal for light field data, as they do not utilize the full potential of light field 4D structure, while existing light field super-resolution algorithms rely heavily on the accuracy of the estimated depth and perform complex sub-pixel disparity estimation. In this paper, we propose a new light field super-resolution algorithm which can address depth uncertainty with a layered object space. First, a pixel-wise depth estimation is performed from the resampled views. Then we divide the depth range into finite layers and back-project pixels onto these layers in order to address the sub-pixel depth error. Finally, two super-resolution schemes: in-depth warping and cross-depth learning, are introduced to super-resolve the views from light field data redundancy. The algorithms is tested with extensive datasets, and the results show that our method attains favorable results in both visual assessment and objective metrics compared to other light field super-resolution methods.

Keywords: Light field; Image warping; Ray-tracing; Super-resolution

1 Introduction

Light field cameras are exploited on the basis of plenoptic theory which was first presented in 1991 by Bergen et al [1]. Later, Levoy et al. [2] developed the theory and parameterized 4D light fields. Light field cameras utilize a special optical design - an array of lenslets inserted between main lens and the sensor, to acquire the ability of capturing an encoded 4D spatio-angular from a single exposure. Instead of integrating incoming rays from a focused scene, the lenslets differentiate the angle information and record it on separate pixels which belong to a same lenslet image. Such plenoptic cameras enable post-processings beyond the capability of conventional cameras, such as rendering refocused and all-in-focus images from computational photography [3, 4]. However, the spatial resolution of the light field cameras is unavoidably reduced significantly due to the spatial-angular tradeoff of the limited number of pixels on the sensor. Such information loss is unbearable for various vision tasks [5, 6] and it is considered as the primary bottleneck for the applications of light field cameras.

The spatial resolution of light field cameras can be improved by two means - optical design and computational photography. The first one is often limited by the wave property of optics such that the size of a pixel cannot go beyond a certain

size, and the high-order optical aberrations come into noticeable effects when the manufacturing requirements is beyond practical limitations. Therefore, the formulation of super-resolution as a computer vision task is of vital importance to make full use of the data captured by light field cameras.

In this work, we propose a depth-assisted light field super-resolution (DASR) algorithm that restores high-frequency details of the views rendered from light field cameras. The algorithm relies on the depth as additional input, and transforms light field spatial super-resolution into a three-dimensional interpolation problem in the object space. The algorithm compromises three steps. The first step is to coarsely estimate the depth of a set of superpixels from multiple views. Based on the depth range, we further divide the object space into multiple layers and remap the 3D objects onto those layers. Finally, a super-resolution filter is applied to the target views based on the re-arranged layers of the light field. The results show that our algorithm compares favorably to state-of-the-art Light Field Super-Resolution (LFSR) algorithms, both in terms of visual quality and objective reconstruction errors. The novelty of this work lies in the three-dimensional super-resolution performed in the object space. It utilizes in-depth warping and cross-depth learning, where in-depth warping implies re-projection of light field data onto several depth layers. The work is inspired by our previous work on depth-based demosaicing [7], but extends the analysis of the layered structure to super-resolve the light field in the object space. In particular, we apply different filters to obtain a super-resolved light field, focusing on the visual quality and quantitative comparison with state-of-the-art super-resolution algorithms of light field.

The article is organized as follows. Section II presents the related works and a super-resolution taxonomy, Section III presents the ray models, sampling analysis of light field capturing systems, and proposed the layered object space as a tool to handle light field. Section IV provides a detailed description of our super-resolution algorithm, followed by experiments and analysis in Section V. Finally, Section VI concludes the article and discuss about the future research interests.

2 Background

2.1 2D image super-resolution

In general, conventional super-resolution can be classified as single-image super-resolution (SISR) and multi-image super-resolution (MISR). In SISR, only one low-resolution view of the scene is employed to obtain a high-resolution output. This is achieved by learning the mapping between low-resolution image and high-resolution, regardless of the training set [8]. Some SISR can learn from the input low-resolution image itself by finding repetitive patterns, some others require an external training set which can provide more abundant high-resolution details [9]. In either way, it migrates the learned information to turn a low-resolution input to a high-resolution masquerade. SISR algorithms can be applied to light field data by taking multiple views of the light field as independent images. However, such favor ignores the correlation among views and fails to exploit inherent details which can be identified when correspondence is considered.

In the multi-image case, the resolution can be enhanced in multiple scales, including both spatial and angular super-resolution. The angular resolution enhancement,

i.e. view synthesis has also been frequently investigated in other works [10,11], and this topic is beyond the scope of this work, instead, we focus on the spatial resolution enhancement. In MISR, a projective transformation which consists of both rotation and translation is often assumed between different views. To perform MISR, such warping is first calculated and then missing pixels are warped from different views. However, the computation of such warping models is demanding, and the geometry structure of light field data does not fit exactly to the warping assumption. Both classic single-image superresoliution and multi-image super-resolution cannot be applied to the light field directly, and the optimal solution need to be tailored to explore correlations amid light field 4D structure.

2.2 Light field spatial super-resolution

The spatial resolution of a light field is not limited by the a single view itself, but the projection from all other views [12–14], hence it is worth noting that additional views are beneficial to spatial resolution enhancement though view synthesis is beyond the scope of this paper [15,16].

By taking advantage of disparity or depth information, a reference view can be super-resolved by propagating pixels that are projected from other views. Projection-based methods is a natural exploration to make use of angular samples of light field, and it focuses on the image formation models of the capturing systems. Lim et al. [12] proposed to consider the sub-pixel disparity which is inherently coded in the angular samples to enhance the resolution of spatial views. Georgiev et al. [13] proposed a similar projection scheme for focused plenoptic cameras, combining super-resolution with the demosaicing process and obtained sharp images with reduced fill rate. In [17], Liang et al. proposed a theoretical ray-tracing model for different capturing systems based on the transfer matrices, and applied inverse light transport to achieve the goal of super-resolution. Rooted from the BM3D filter for image denoising [18], Alain and Smolic [19] proposed an LFBM5D that is tailored for 4D light field. The algorithm iteratively alternates between LFBM5D filtering and back-projection for LFSR.

Optimization-based methods solve the super-resolution problem with different models and priors. [20] introduced a variational Bayesian framework to super-resolve the light field by merging multi-view information. The Lambertian reflectance and texture-preserving priors are employed to avoid aliasing. Mitra et al. [21] proposed a patch-based approach using a Gaussian mixture model (GMM) such that each patch is considered as a Gaussian random variable conditioned on its disparity. The GMM patch prior is then used to perform multiple tasks, including super-resolution. Wanner et al. [22] employed a total variation prior not only to obtain a favorable inpainting result, but also to minimize computational effort in the variational framework. The framework can achieve both spatial super-resolution and angular view synthesis. An undirected weighted graph model was constructed in [23] with a HR-LR correlation term, a view-based constraint and a high-resolution geometric structure regularization respectively.. The work is further extended in [24] and [25] to replace the quadratic regularizer (which tend to be low-pass) with a non-smooth regularizer in order to preserve high frequency information and to reduce the computational complexity by ignoring diagonal views, respectively.

More recent advancement in LF spatial super-resolution are mostly based on learning methods due to its outstanding performance when huge data are available. Driven by vast data, CNN was also introduced to light field super-resolution in [26], where two networks are trained for spatial super-resolution and angular view synthesis independently. Gul et al. proposed to collect four neighboring pixels from an SAI and use a shallow CNN to predict three in-between super-resolved pixels to achieve a higher spatial resolution [27]. A two-step approach is proposed in [28] where geometric structure of views is used upon individual views and then a trained network is applied to enforce correct parallax information. Another similar work is proposed in [29] where the super-resolution problem is decomposed into three steps - initialization, view alignment and cross-view correlation, each solved with a specialized CNN.

Unlike optimization-based and learning-based methods which heavily rely on computational power and the abundant training procedure, the projection-based methods take advantage of both geometric structure of the scene and image formation models. The work in this paper can be categorized as projection-based Super-Resolution (SR) method. However, different from requirements of sub-pixel accuracy of depth estimation, we try to solve the problem of super-resolution problem under the existence of uncertain depths. Furthermore, A combination of in-depth warping and cross-depth learning are proposed in a layered object space to facilitate the super-resolution process during the back-projection.

3 Methods

In this section, we propose a layered object space architecture that we used to super-resolve the general light field data. We first discuss about different capturing setups, their parameterizations and how they can be integrated into general light field representation. Then we introduce the layered object space where we carry out super-resolution.

3.1 Light field parameterization

Light field can be acquired using different devices, including single exposure of plenoptic camera, multiple exposures of a static scene by a moving camera gantry, and synchronized captures of a camera-array towards a moving object, thus bringing various parameterizations to the light field data.

In a conventional plenoptic camera, each microlens gathers angular information converged by the main lens. In spatial domain, the sampling distance p in the object space is proportional to the microlens pitch μ. Additionally, the spatial resolution of the scene is equivalent to the number of microlenses N. A subaperture image can be rendered by collecting corresponding pixels under each microlens, representing the scene from one angle of view. However, p becomes irregular in the object space when the scene is off the main lens focal plane, shown as p' in Fig. 1. In this case, each microlens samples a small region of the scene instead of a singular spatial point. The maximal sampling frequency depends on the projection of non-adjacent pixels AC', rather than that of adjacent sensor pixels AB. Different from plenoptic camera, in multi-camera system, such as camera gantry or camera array, the spatial and angular sampling frequency is limited by the camera resolution and the number of exposures, respectively.

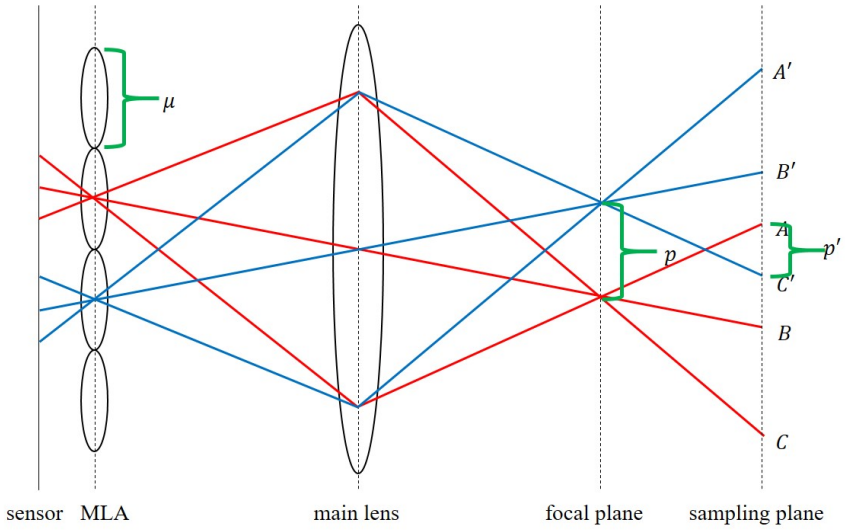


Figure 1: A ray model of plenoptic cameras.

In the light field literature, we note that it is common to use the general definition of light field to include all of the aforementioned parameterizations. The light field parameterizes light rays of the scene with the coordinates of their intersection withtwo parallel planes. Each light ray is associated to a radiance value, and located by the coordinates of intersections on these two planes. This is the sampling scheme approximated by both camera arrays and light field cameras. In the following, the light field is defined as an $M \times M$ array of virtual pinhole cameras, each one equipped with an $N \times N$ pixel sensor. The coordinates of camera centers are defined on a virtual camera plane, with its coordinates (s, t). Different cameras (s, t) essentially represent different views, thus it is hereafter referred as the angular coordinates. Similarly, the pixel coordinates (x, y) are defined on the sensor plane, representing the spatial information recorded by a view. The baseline B defines the distance between neighboring virtual cameras on the $s - t$ plane, and the distance between the $s - t$ plane and the $x - y$ plane is called the focal length f of the light field capturing system. As same to the convention, the sub-aperture views of light field cameras or the alignment of conventional cameras of the light field capturing systems are considered rectified.

Consider a point P at depth z from a view, and its projection on one of the cameras $U_{s,t}$ is represented by the pixel $U_{s,t}(x, y)$. The projection of P on the other views are $U_{s\prime,t}(x\prime, y)$, in the same row of the apertures, assuming that only horizontal parallax is applied. Thus, t and y will retain the same. When there is no occlusion, we can refer that

$$U_{s,t}(x, y) = U_{s\prime,t}(x\prime, y) \tag{1}$$

under the Lambertian assumption.

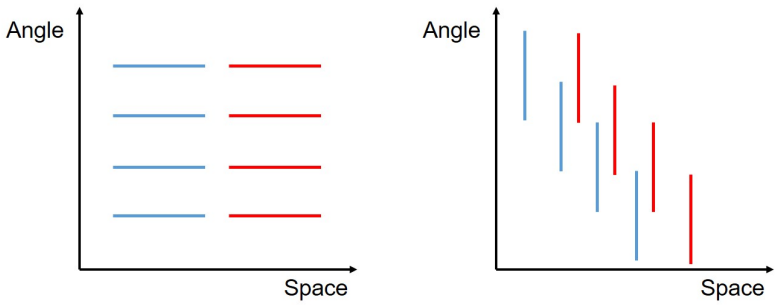


Figure 2: Phase space representation of light field camera sampling. Left: plenoptic 1.0; right: plenoptic 2.0.

3.2 Layered object space

For simplicity and consistency of the paper, the following analysis is based on the sampling model of plenoptic cameras (Fig. 1), similar analysis can be performed for multi-camera setup, as discussed in the previous section. We can see from Fig. 1 that on a sampling plane, the two neighboring pixels behind the same microlens yield a sampling distance AB which is greater than the sampling distance $AC\prime$ when pixels of other microlenses are taken into account. This means that applying Shannon's Theorem separately to each microlens which disregards the Micro-Lens Array (MLA) structure will produce chromatic artifacts and erroneous interpolation results. A direct solution is to apply single-image super-resolution to individual views $U_{s,t}$. In essence, such an SISR uses pixels at the same relative lenslet positions (x, y) for LFSR, regardless of correspondence pairs as in 1. This causes a waste of the captured information as it does not consider the full light field.

We can draw a same conclusion from phase space sampling diagrams and ray tracing. As shown in Fig. 2, for plenoptic 1.0, the ray transfer matrix is:

$$\mathbf{A} = \begin{bmatrix} 1 & f \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} = \begin{bmatrix} 0 & f \\ -\frac{1}{f} & 1 \end{bmatrix}, \tag{2}$$

where $\mathbf{A}$ is the ray transfer matrix, and $\mathbf{A^{-1}} = \begin{bmatrix} 1 & -f \\ \frac{1}{f} & 0 \end{bmatrix}$ indicates the phase space changes from main lens to the microlens sampling. Therefore, the sampling of plenoptic 1.0 is performed as a rotation of each pixel to 90 degrees in optical phase space, and it causes the low spatial resolution of such design.

On the other hand, for plenoptic 2.0, the transfer matrix is:

$$\mathbf{A} = \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -\frac{b}{a} & 0 \\ -\frac{1}{f} & -\frac{a}{b} \end{bmatrix}, \tag{3}$$

where a and b are the distances from the object and its image with respect to the main lens. The inverse matrix $\mathbf{A^{-1}} = \begin{bmatrix} -\frac{a}{b} & 0 \\ \frac{1}{f} & -\frac{b}{a} \end{bmatrix}$ indicates that there is no rotation

of phase space, but only shearing, as shown in the right of Fig. 2, which means that the sampling of plenoptic 2.0 is more flexible as it decouples spatial resolution from number of microlenses and allows to choose the spatial-angular tradeoff point freely.

Both ray model (Fig. 1) and phase space analysis (Fig. 2) show that to solve the LFSR problem, one must consider the contributions of other microlenses. Thus, projection-based LFSR methods calculate the sub-pixel shift based on the camera parameters and correspondence searching methods. Unfortunately, disparity estimation is still not fully solved, and mostly suffers from heavy computational burden. Furthermore, the performance of such disparity estimation is compromised when the sampling is too sparse to search for accurate correspondences.

Consider the ray model in Fig. 1, the minimal sampling distance varies from p to $p\prime$ based on the location of the object, and the nearest sample is not always captured by the adjacent microlens. This implies that either sub-pixel interpolation is needed when searching for correspondences across the views, or Eq. 1 is compromised by finding suboptimal matches and enforcing such photo-consistency from available pixels, both resulting in an erroneous disparity.

Generally, enforcing the photo-consistency constraint is carried out by finding a similarity score among views. We define the similarity score between the target pixel $U_{s,t}(x,y)$ and a generic pixel $U_{s\prime,t\prime}(x\prime,y\prime)$ as follows:

$$\omega_{s\prime,t\prime}(x\prime,y\prime) = \exp\left(-\|\mathcal{P}_{s\prime,t\prime}(x\prime,y\prime) - \mathcal{P}_{s,t}(x,y)\|_F^2\right), \tag{4}$$

where $\mathcal{P}_{s,t}(x,y)$ is a square patch centered at the pixel $U_{s,t}(x,y)$ with $W \times W$ as its window size, and the operator $\|\cdot\|_F$ is the Frobenius norm. Unlike other pixel-wise correspondence method which need to set a 'relaxed' threshold of similarity in order to fulfill consistency constraints in all neighboring views $s\prime = \{s \pm 1, s \pm 2...\}, t\prime = \{t \pm 1, t \pm 2...\}$, we set a relatively hard threshold to search the correspondence from both adjacent and distance views. Thus, W is the disparity range variable that calculated for each pair of views according to their relative camera position. In other words, we constrain the photo-consistency strictly so that only robust matches are registered, and the correspondences are not enforced for all the view pairs. In this way, the correspondence searching process finds stable matches across a relatively large baseline.

We further note that the surface radiance of a natural object tends to be smooth, the scattered 3D samples in the object space can be assigned to several surfaces depending on this photo-consistency measure. The depth range is controlled by the disparity range: $z = \frac{fB}{d}$, where $\mathcal{Z}$ is the depth, and B is the baseline between $U_{s,t}$ and $U_{s\prime,t}$. We back-project $U_{s,t}(x,y)$ according to the estimated disparity to generate a group of scattered 3D samples in the object space, as summarized in Algorithm 1.

The layered object space keeps the physical sampling relationships among pixels, and such information is used to boost the reconstruction of 3D information. Note that till now the pixels $U_{s,t}(x,y)$ from all views are re-arranged in planar manners on separated depth layers.

Algorithm 1 Generation of layered 3D object space samples

Input: Light field L, Bins of depth layers $(min_1, max_1), \dots (min_k, max_k)$
Output: 3D samples in the layered object space $U(\tilde{x}, \tilde{y}, \tilde{z})$

```
for each view (s, t) of light field do
    Build ray transfer matrix A
    Calculate back-projection matrix A^{-1}
    for each pixel U_{s,t}(x, y) do
        Estimate depth ẑ for pixel U_{s,t}(x, y)
        Adsorb ẑ to its nearest layer z̃
        Back-projection U(x̃, ỹ, z̃)) = A^{-1} U_{s,t}(x, y)
        Register U(x̃, ỹ, z̃) in layered space
    end
end
```

4 Super-resolution algorithm

4.1 Problem formulation

In general, to perform super-resolution in 3D space is a complicated problem due to the non-uniform sampling on different depths. However, there exists extreme cases which are simple to cope with: 1) when the scene is entirely on the focal plane, and 2) when the planar scene is off on the focal plane, and perpendicular to the optical axis of the camera, as shown in Fig. 1. The first Case can be deemed as a special case of case 2) that conventional SISR can be directly applied. In the second case, we introduce our LFSR algorithm which consists of an in-depth and a cross-depth superresolution. Note that even though the scene is planar, the minimum sampling distance varies based on the depth. Therefore, an inverse distance weighting (IDW) function is adopted to generate regular grid on each layer:

$$U(\tilde{x}, \tilde{y}, \tilde{z}) = \sum_{i=1}^{m} \omega_i U(x, y, z) \Big/ \sum_{i=1}^{m} \omega_i \ , \tag{5}$$

where $\omega_i = (d_i)^{-n}$ is the weighted distance between 3D sample (x, y, z) and its i-th nearest neighbor, and n defines the decaying of weighting with the distance d_i which affects the smoothness of the pixel grid on each depth layer $\tilde{z}$.

4.2 In-depth warping

Classic image super-resolution often relies on the powerful image priors such as patch recurrence and edge structure. These explicit priors form the basis of both projection-based and optimization-based methods. In this section, we adapt the notion of patch recurrence to the 4D light field.

In MISR, a warping function is used to stand for any transformation between the common high-resolution source H and a set of its observed low-resolution images L. With the inclusion of a blurring kernel and the subsampling process, a simple warping function, if no rotation or translation is considered, can be written as:

$$L_j = \mathbf{S_j B_j} H, \tag{6}$$

where S_j and B_j denotes the subsampling operation and the corresponding blurring kernel in the matrix form. Each low-resolution image L_j can give a rise to finding the high-resolution image H. If enough low-resolution images are observed (determined

by the support of the subsampling factor), H can be determined. Thus, the super-resolution problem is formulated as an inverse problem of the warping process, and it can be optimized when over-determined.

In principle, the problem of recovering the mutual high-resolution image H for the light field needs a different warping function from Eq. 6, as the rotation and translation between views should be considered. If the light field capturing system is calibrated, i.e. knowing the projective transformation matrix between the scene content and the image, then a fundamental matrix can be uniquely determined. However, the fundamental matrix (the algebraic representation of epipolar geometry) only represents the singular correlation between a point from one image to its corresponding epipolar line in the second image. To find pixel correspondences, depth information is needed as another degree of freedom to the matrix.

Fortunately, depth can be inferred from the multi-view stereo of a light field by depth estimation algorithms. Furthermore, by using the ray-tracing technique in Section 3.2, one can project pixels from views to the object space following its optical path, integrating light rays onto several depth layers. Thanks to the abundant angular information of light field, repetitive patches can be found with small effort on each depth layer.

As each depth layer corresponds to a specific disparity, after the first patch of U_{s_i,t_i} is located, its $M \times M$ similar patches captured by any other view U_{s_j,t_j} can be found by shifting the patch according to the vector $\mathbf{v_{i,j}}$:

$$\mathbf{v_{i,j}} = \left(\frac{(s_j - s_i)fB}{z}, \frac{(t_j - t_i)fB}{z} \right)^T . \tag{7}$$

By finding nearest patch neighbors [30] on each depth layer, a light field super-resolution problem is reformulated as an integration of a set of classic MISR problems, and the MISR image warping model can be solved independently on each layer. The idea can be summarized to the following simple algorithm: For each pixel on a low-resolution view image, find its depth and project it to the nearest depth layer following its optical path. Each depth layer gather information contributed by multiple views of the same scene, and patch recurrences are used to estimate the warping function 6. If enough patch recurrences are found, the super-resolution can be performed successfully on the depth layer.

On the other hand, if the similar patches are not found from other views on the same projected depth layer, some other schemes need to be used to recover high-resolution image. In fact, this happens if only the depth is erroneous and the corresponding patches are projected onto different layers. In this case, further error propagation is stopped. In essence, the number of layers essentially set the trade-off between accurate patch correlation and erroneous depths. The more accurate a depth map is, the more layers should be adopted to ensure a high patch similarity score. This means that high-resolution image H can be recovered in a well-defined manner. The more defective a depth map is, the less layers should be employed to enforce more matching patches and make the algorithm robust to minor depth errors.

4.3 Cross-depth learning

The in-depth warping extends the classic MISR algorithm to be applicable to the 4D light field data. However, due to the erroneous depth estimation process and possible additive noise, the depth-assisted light field super-resolution still suffers from insufficient patch recurrences.

The learning technique has developed rapidly over the last few years and it has been shown to exceed the limits of classical SR. However, the essence remains the same: to find a mapping from the input to the output. In this section, we show how similar ideas can be exploited in our LFSR framework without any external datasets for training. It utilizes a simple and efficient learning process by employing patch repetitions during the depth layer generation described in Algorithm 1.

Let $\{z\prime\} \in \mathcal{Z}$ be the depth layer that needs to be super-resolved by a factor α. As depth layers moves away from the camera center, the projection of sensor onto the layer will be enlarged by a factor z/f. This means that the closer a depth layer is, the more densely a scene is sampled by the light field. Let $\Pi = \tilde{z}_0, \tilde{z}_1, \tilde{z}_2, ...$ denote a cascade of depth layers of increasing resolution so that larger index indicates a smaller distance to the camera. Therefore, we can search for similar patches within the high resolution layer z/α, using approximate nearest neighbor search [30] in different depth layers (scales). This provides a high- and low-resolution patch pair that generated from the same light field. Once an approximate nearest neighbor of the target low-resolution patch is found on a lower-scale, the high-resolution mapping of the nearest neighbor on a different depth layer is migrated to super-resolve the target patch. One should note that like any other learning-based methods, cross-depth learning migrate high-resolution details from its similar patch from another scene structure, which means that the details are hallucinated. Therefore, the in-depth warping plays the central role, whereas the cross-depth learning is a supplement to in-depth warping.

5 Results and discussion

5.1 Experimental setup

Table 1: Evaluated Super-resolution Methods for Light Field Data

Method	Language	Category	Time (s)
Bicubic	Matlab	Single image	0.22
LFBM5D [19]	Matlab & C++	Projection-based	601.17
GB [24]	Matlab	Optimization-based	9×10^5
LFCNN [27]	Python	Learning-based	68.93
Proposed	Matlab	Projection-based	170.19

We consider both analytic interpolation (bicubic interpolation) and light field super-resolution algorithms in our experiments. We carefully choose projection-based LFBM5D [19], optimization-based GB [24], and learning-based LFCNN [27] to compare with our results, each of which belongs to a different category from our taxonomy in Section 2.2, and shows the state-of-the-art progress in light field super-resolution. Table 1 summarizes the details about different comparison methods, including implementation language, category, and average computational time (using the same size of cropped datasets to perform super-resolution). Our experimental setup consists of an Intel Core i5 650 dual core CPU with 12GB RAM,

Table 2: Mean PSNR and SSIM for the SR factor $\alpha = 2$

	HCI		Stanford	
	PSNR	SSIM	PSNR	SSIM
Bicubic	29.14	0.900	35.93	0.911
LFBM5D [19]	33.58	0.954	32.72	0.943
GB [24]	35.47	0.971	39.33	0.959
LFCNN [27]	37.21	0.979	41.13	0.976
Proposed	36.79	0.970	41.01	0.964

and the running time is measured based on the CPU implementation of respective methods.

We conduct our experimental comparison with two publicly available datasets: the HCI light field dataset [31] and the (New) Stanford light field dataset [32]. The HCI dataset is consisted of 12 synthetic light fields, and the angular resolution is 9×9. The light fields of HCI dataset has a relatively small baseline, resulting in a disparity range within $[-3, 3]$, simulating the light field camera data representation. On the contrary, the Stanford dataset is captured by a Lego camera gantry, consisting of real static scenes captured by a movable conventional camera with a large baseline distance. The typical angular resolution of the Stanford dataset is 17×17, and each view has a high spatial resolution which is captured by the full camera sensor. Therefore, the Stanford dataset resembles both camera gantry and camera array scenario well. Furthermore, the Stanford dataset suffers from vignetting effects and optical aberrations. Therefore, similarly to experiments conducted in other super-resolution works such as [19], we crop the Stanford light fields to a 5×5 array of views, i.e. $M = 5$, to minimize the effects of image degradations and speed up the experiments.

In our experiments, the spatial resolution of each light field U is first decreased by a factor $\alpha = 2$ by applying a subsampling matrix and a blurring filter to each view. Then, the low resolution light field is super-resolved by the same factor $\alpha = 2$ in order to compare with the original image. Such procedure is repeated for all the studied algorithms in this section.

5.2 Results and analysis

The objective quality metrics of the studied super-resolution algorithms are shown in Table 2, with a super-resolution factor $\alpha = 2$. For each super-resolved light field we compute the Peak Signal-to-Noise Ratio (PSNR [dB]) and Structural Similarity Index (SSIM) at each view and report the average of the computed PSNR and SSIM in Table 2. PSNR reflects the fidelity of the signal, and SSIM is a complementary measure that takes perception of human visual systems into account.

As can be seen from Table 1 and Table 2, conventional SR methods like LFBM5D and GB suffer from heavy computational load, while standard bicubic interpolation performs simple but inefficient interpolation. Our method performs well in PSNR (second best for both datasets) while saving computational effort. Though the execution time of CNN-based methods is short, it requires a heavy training beforehand. The advantages of the proposed DASR algorithm in computational efficiency will be more significant if one considers the parallelized programming for super-resolution of multiple views of the light field.

Table 3: Comparison between different super-resolution methodologies (SR factor $\alpha = 2$) with PSNR metric using HCI dataset

	Bilinear [33]	[19]	[24]	[21]	[34]	[35]	[36]	Proposed
buddha	35.22	33.01	39.00	39.12	37.73	38.42	39.11	**39.17**
buddha2	30.97	31.79	34.41	33.63	33.67	35.4	36.04	**37.11**
couple	25.52	34.55	33.51	31.83	28.56	33.31	33.91	**36.09**
cube	26.06	30.68	33.28	30.99	28.81	33.18	33.78	**34.51**
horses	26.37	31.86	32.62	33.13	27.80	33.02	33.62	**34.63**
maria	32.84	34.91	37.25	37.03	35.50	38.32	**39.02**	37.64
medieval	30.07	33.49	33.45	33.34	31.23	33.40	35.12	**35.41**
mona	35.11	36.76	39.37	38.32	39.07	38.73	**40.72**	39.73
papillon	36.19	34.16	40.70	40.59	39.88	40.65	**42.74**	39.44
statue	26.32	35.10	35.61	32.95	29.65	33.97	35.72	**36.19**
stillLife	25.28	33.11	30.98	28.84	27.27	30.14	31.69	**34.77**

Table 4: Comparison between different super-resolution methodologies (SR factor $\alpha = 2$) with PSNR metric using Stanford dataset

	Bilinear [33]	[19]	[24]	[21]	[34]	[35]	[36]	Proposed
amethyst	35.59	28.17	39.19	36.08	38.81	39.51	39.3	**39.79**
beans	47.92	40.92	48.41	36.02	52.01	**54.68**	50.7	49.16
bracelet	33.02	29.67	28.27	19.91	38.05	**44.37**	28.6	32.96
bulldozer	34.94	34.15	35.96	32.05	39.76	**45.79**	36.5	41.77
bunny	42.44	36.91	47.01	40.66	47.16	48.01	48.45	**49.04**
cards	29.50	30.41	36.52	37.03	33.77	36.45	38.25	**39.37**
chess	36.36	34.49	41.86	34.74	40.75	43.58	**43.95**	40.06
eucalyptus	34.09	31.55	39.09	34.90	36.69	37.35	40.43	**41.47**
knights	34.31	31.88	37.23	29.33	38.37	**39.11**	39.10	38.92
treasure	30.83	28.19	37.51	30.52	34.16	34.77	37.62	**38.51**
truck	36.26	33.56	41.57	37.52	39.11	39.81	**42.52**	40.16

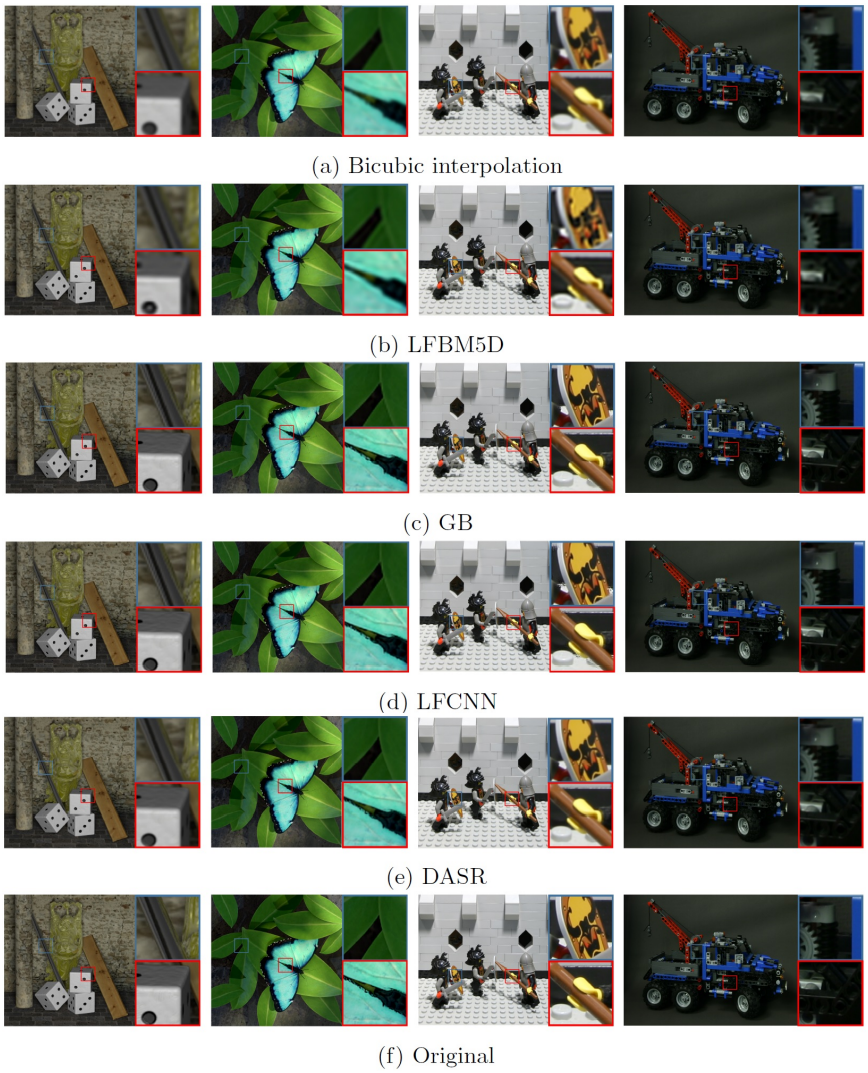
It can be seen from Table 2 that the proposed DASR outperforms other super-resolution algorithms except LFCNN. It is worth pointing out that both DASR and LFCNN achieve high-fidelity results and visual difference can hardly be identified in subjective tests (Fig. 3). However, deep learning methods require massive external datasets and a considerable training time before a swarm of parameters can be well-tuned to perform super-resolution tasks. On the other hand, optimization-based methods (GB) is extremely demanding on the computational time to achieve comparable objective results with the proposed DASR due to its iterative optimization. Furthermore, GB has a drastic drop in its performance with Stanford dataset. This is due to the unexpected large disparity range (resulted from multi-camera large baseline) which GB cannot handle. The LFBM5D is relatively fast compared with GB, but it fails to recover fine details of the original scene, as the main goal of such a method is to remove noises from images. The effectiveness of the proposed method is further validated by comparing PSNR results of each light field using different light field super-resolution methods in Table 3 and Table 4. As can be seen, the proposed method has the most favorable outcome regarding reconstruction error in 13 out of 22 light fields in both databases.

Fig. 3 shows the visual comparison of different super-resolution methods. The center views of four scenes from HCI and Stanford dataset are chosen for comparison. As can be seen, DASR recovers fine details across depth discontinuities with the help of the layered object space structure. In general, GB, LFCNN, and DASR render more desirable results than bicubic interpolation and LFBM5D. The visual difference between GB, LFCNN and DASR can hardly be perceived by human eyes, and this is achieved by DASR without heavy computational power or pre-trainings, which demonstrates the effectiveness of the in-depth warping and cross-depth learning schemes. Note that even though LFCNN outperforms DASR in PSNR, it creates unrealistic scene structure in the super-resolution results (see scene *Knights*), and this is due to the fact that LFCNN learns the scene structure from external datasets, which unavoidably leads to an unreliable detail transfer from irrelevant scene features.

6 Conclusions

In this paper, we proposed a light field super-resolution algorithm in the layered object space. Thanks to the in-depth patch warping, we circumvent the intractable problem of estimating a global image warping model without requiring external datasets for the training process. Cross-depth patch learning further enables a stable warping model when the angular resolution is low. The proposed algorithm compares favorably to the state-of-the- art light field super-resolution frameworks, both in terms of visual quality and in terms of quantitative assessments. Additionally, as the proposed method super-resolve the light field in the object space, it is suitable to handle light field data irrespective of its capturing method and disparity range.

Despite the spatial super-resolution discussed above, when the scene is not considered as Lambertian, one can develop sophisticated algorithm based on the angular sampling information $U(\theta, \varphi, z)$ rather than spatial sampling information $U(x, y, z)$. The conversion between angular coordinates and spatial coordinates is:



(a) Bicubic interpolation

(b) LFBM5D

(c) GB

(d) LFCNN

(e) DASR

(f) Original

Figure 3: Visual results of different SR methods on Buddha, Papillon, Knights and Truck with a scaling factor $\alpha = 2$.

$\theta = \arctan\frac{x-s}{f}, \varphi = \arctan\frac{y-t}{f}$, where θ and φ are the horizontal and vertical angles respectively. If the scene is of complex lighting conditions, i.e. non-Lambertian, the differences in θ and φ can be integrated into Eq. 4 as complementary terms.

Abbreviations

LF: Light field; 2D: Two-dimensional; 3D: Three-dimensional; 4D: Four-dimensional; SR: Super-resolution; DASR: Depth-assisted light field super-resolution; SISR: Single-image super-resolution; MISR: Multi-image super-resolution; GMM: Gaussian mixture model; CNN: Convolutional neural network; MLA: Micro-lens array; PSNR: Peak signal-to-noise ratio; SSIM: structural similarity index measure

Acknowledgements

Not applicable.

Authors' contributions

YL did the main work, implemented the proposed method and carried out the experiments. MS participated in data analysis and discussion. Both authors took part in the writing and proof reading of the final version of the paper. The authors read and approved the final manuscript.

Authors' information
Yongwei Li received the B.Sc. and M.Sc. degrees in Computer Science and Technology from Liaoning Normal University, China in 2012 and 2015 respectively, and the PhD degree from the Department of Information Systems and Technology (IST), Mid Sweden University, Sweden in 2020. His research interests are image processing and light field imaging.
Mårten Sjöström received the M.Sc. degree in electrical engineering and applied physics from Linköping University, Sweden, in 1992, the Licentiate of Technology degree in signal processing from the Royal Institute of Technology, Stockholm, Sweden, in 1998, and the Ph.D. degree in modeling of nonlinear systems from the École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2001. He was an Electrical Engineer with ABB, Sweden, from 1993 to 1994, a fellow with CERN from 1994 to 1996, and a Ph.D. Student at EPFL, Lausanne, Switzerland, from 1997 to 2001. In 2001, he joined Mid Sweden University, and he was appointed as an Associate Professor and a Full Professor of Signal Processing in 2008 and 2013, respectively. He is the head of subject Computer and System Sciences since 2013 and of Computer Engineering since 2020. He founded the Realistic 3D Research Group in 2007. His current research interests are within multidimensional signal processing and imaging, and system modeling and identification.

Funding
The work in this paper was funded in part from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 676401, European Training Network on Full Parallax Imaging.

Availability of data and materials
The image datasets used to support the findings of this study can be downloaded from the public websites whose references are provided in the article.

Competing interests
The authors declare that they have no competing interests.

Author details
Department of Information Systems and Technology, Mid Sweden University, Sundsvall, Sweden.

References

1. Bergen, J.R., Adelson, E.H.: The plenoptic function and the elements of early vision. Computational Models of Visual Processing, 3–20 (1991)
2. Levoy, M., Hanrahan, P.: Light field rendering. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 31–42 (1996). ACM
3. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Computer Science Technical Report CSTR **2**(11), 1–11 (2005)
4. Peng, L., Dijun, L.: All-in-focus image reconstruction based on plenoptic cameras. In: 2013 Seventh International Conference on Image and Graphics, pp. 612–617 (2013). IEEE
5. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
6. Oh, S.J., Benenson, R., Khoreva, A., Akata, Z., Fritz, M., Schiele, B.: Exploiting saliency for object segmentation from image level labels. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5038–5047 (2017). IEEE
7. Li, Y., Sjöström, M.: Depth-assisted demosaicing for light field data in layered object space. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3746–3750 (2019). IEEE
8. Nasrollahi, K., Moeslund, T.B.: Super-resolution: a comprehensive survey. Machine vision and applications **25**(6), 1423–1468 (2014)
9. Timofte, R., Rothe, R., Van Gool, L.: Seven ways to improve example-based single image super resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1865–1873 (2016)
10. Kalantari, N.K., Wang, T.-C., Ramamoorthi, R.: Learning-based view synthesis for light field cameras. ACM Transactions on Graphics (TOG) **35**(6), 1–10 (2016)
11. Wang, Y., Liu, F., Wang, Z., Hou, G., Sun, Z., Tan, T.: End-to-end view synthesis for light field imaging with pseudo 4dcnn. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 333–348 (2018)
12. Lim, J., Ok, H., Park, B., Kang, J., Lee, S.: Improving the spatail resolution based on 4d light field data. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 1173–1176 (2009). IEEE
13. Georgiev, T., Chunev, G., Lumsdaine, A.: Superresolution with the focused plenoptic camera. In: Computational Imaging IX, vol. 7873, p. 78730 (2011). International Society for Optics and Photonics
14. Damghanian, M., Olsson, R., Sjöström, M., Fructuoso, H.N., Martinez-Corral, M.: Investigating the lateral resolution in a plenoptic capturing system using the spc model. In: Digital Photography IX, vol. 8660, p. 86600 (2013). International Society for Optics and Photonics
15. Wang, Y., Wang, L., Yang, J., An, W., Yu, J., Guo, Y.: Spatial-angular interaction for light field image super-resolution. In: European Conference on Computer Vision, pp. 290–308 (2020). Springer
16. Ivan, A., Park, I.K., *et al.*: Joint light field spatial and angular super-resolution from a single image. IEEE Access **8**, 112562–112573 (2020)
17. Liang, C.-K., Ramamoorthi, R.: A light transport framework for lenslet light field cameras. ACM Transactions on Graphics (TOG) **34**(2), 1–19 (2015)
18. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on image processing **16**(8), 2080–2095 (2007)
19. Alain, M., Smolic, A.: Light field super-resolution via lfbm5d sparse coding. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 2501–2505 (2018). IEEE

20. Bishop, T.E., Favaro, P.: The light field camera: Extended depth of field, aliasing, and superresolution. IEEE transactions on pattern analysis and machine intelligence **34**(5), 972–986 (2011)
21. Mitra, K., Veeraraghavan, A.: Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 22–28 (2012). IEEE
22. Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. IEEE transactions on pattern analysis and machine intelligence **36**(3), 606–619 (2013)
23. Rossi, M., Frossard, P.: Graph-based light field super-resolution. In: 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6 (2017). IEEE
24. Rossi, M., El Gheche, M., Frossard, P.: A nonsmooth graph-based approach to light field super-resolution. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 2590–2594 (2018). IEEE
25. Yim, J., Van Duong, V., Jeon, B.: Time complexity reduction on light-field super-resolution with graph-based regularization. In: International Workshop on Advanced Imaging Technology (IWAIT) 2021, vol. 11766, p. 117662 (2021). International Society for Optics and Photonics
26. Yoon, Y., Jeon, H.-G., Yoo, D., Lee, J.-Y., So Kweon, I.: Learning a deep convolutional network for light-field image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 24–32 (2015)
27. Gul, M.S.K., Gunturk, B.K.: Spatial and angular resolution enhancement of light fields using convolutional neural networks. IEEE Transactions on Image Processing **27**(5), 2146–2159 (2018)
28. Jin, J., Hou, J., Chen, J., Kwong, S.: Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2260–2269 (2020)
29. Cheng, Z., Xiong, Z., Chen, C., Liu, D., Zha, Z.-J.: Light field super-resolution with zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10010–10019 (2021)
30. Arya, S., Mount, D.M.: Approximate nearest neighbor queries in fixed dimensions. In: SODA, vol. 93, pp. 271–280 (1993)
31. Wanner, S., Meister, S., Goldluecke, B.: Datasets and benchmarks for densely sampled 4d light fields. In: VMV, vol. 13, pp. 225–226 (2013). Citeseer
32. Laboratory, T.C.G.: The (New) Stanford Light Field Archive. `http://lightfield.stanford.edu/index.html`. [Online; accessed 10-October-2017] (2008)
33. Gonzalez, R.C., Woods, R.E., et al.: Digital image processing. Prentice hall Upper Saddle River, NJ (2002)
34. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European Conference on Computer Vision, pp. 184–199 (2014). Springer
35. Wang, Y., Liu, F., Zhang, K., Hou, G., Sun, Z., Tan, T.: Lfnet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution. IEEE Transactions on Image Processing **27**(9), 4274–4286 (2018)
36. Ghassab, V.K., Bouguila, N.: Light field super-resolution using edge-preserved graph-based regularization. IEEE Transactions on Multimedia **22**(6), 1447–1457 (2019)