A Generative Model for Generic Light Field Reconstruction

Paramanand Chandramouli*, Kanchana Vaishnavi Gandikota*, Andreas Goerlitz,
Andreas Kolb, Michael Moeller
E-mail: {paramanand.chandramouli, kanchana.gandikota, andreas.goerlitz, andreas.kolb, michael.moeller}@uni-siegen.de

Abstract—Recently deep generative models have achieved impressive progress in modeling the distribution of training data. In this work, we present for the first time a generative model for 4D light field patches using variational autoencoders to capture the data distribution of light field patches. We develop a generative model conditioned on the central view of the light field and incorporate this as a prior in an energy minimization framework to address diverse light field reconstruction tasks. While pure learning-based approaches do achieve excellent results on each instance of such a problem, their applicability is limited to the specific observation model they have been trained on. On the contrary, our trained light field generative model can be incorporated as a prior into any model-based optimization approach and therefore extend to diverse reconstruction tasks including light field view synthesis, spatial-angular super resolution and reconstruction from coded projections. Our proposed method demonstrates good reconstruction, with performance approaching end-to-end trained networks, while outperforming traditional model-based approaches on both synthetic and real scenes. Furthermore, we show that our approach enables reliable light field recovery despite distortions in the input.

1 Introduction

High quality light field (LF) images are vital for a wide range of applications such as the precise free viewpoint rendering of a 3D scene or the estimation of geometries or materials of objects in a scene. Mathematically, light fields are represented using the plenoptic function that models the radiance of the scene in spatial and angular dimensions. Unfortunately, the acquisition of high quality light field data is commonly restricted by specific constraints imposed by the underlying camera hardware. Light field images can be acquired using exhaustive and expensive hardware setups comprising dozens of cameras in a camera-rig, or by using *plenoptic cameras* that utilize microlens arrays placed in front of the imager of a standard 2D camera [1]. While camera-rigs allow for larger baselines with rather sparse angular resolution, plenoptic cameras allow recording dense light fields with a rather small baseline. Plenoptic cameras have the advantage that they capture a full light field with a single exposure, but there is a trade-off between the spatial resolution of each sub-aperture image and the angular resolution of the light field.

To address the trade-off between spatial and angular resolution optimally, researchers have proposed to linearly compress the angular or spatial dimension (or both), giving rise to the important problem of recovering a light field $\mathbf{l}$ from linear observations $\mathbf{i}$ related via

$$\mathbf{i} = \mathbf{\Phi}\mathbf{l} + \mathbf{n}, \tag{1}$$

for a (problem dependent) linear operator $\mathbf{\Phi}$ and additive noise $\mathbf{n}$.

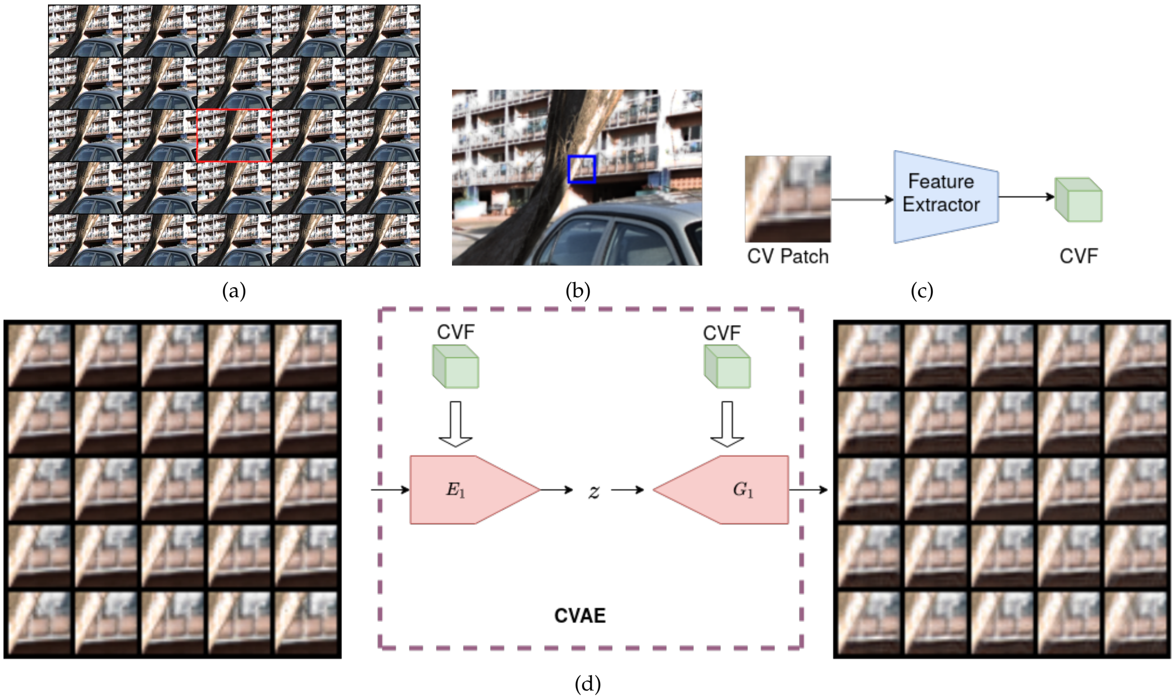
A classical approach to solve the ill-posed inverse problem (1) is by *energy minimization methods*. One designs a cost function H depending on the light field in such a way that low values of $H(\mathbf{l})$ correspond to light fields $\mathbf{l}$ with desirable properties. Subsequently, the solution is determined by finding the argument that minimizes the energy H, for example [2]. An alternate traditional approach is to estimate parameters such as depth map or disparity map which are subsequently used to synthesize light field [3].

Recent approaches have instead simulated large numbers of pairs $(\mathbf{i}, \mathbf{l})$ and learned a mapping from $\mathbf{i}$ to $\mathbf{l}$ by a deep neural network, see [4], [5], [6], [7], [8]. While such approaches often improve the reconstruction quality in a specific application significantly, they lack the flexibility of classical methods and have to be retrained as soon as the observation model (1) changes.

To exploit the expressive power of neural networks without loosing the flexibility of energy minimization methods several hybrid methods have been proposed, e.g. by using neural networks as proximal operators (often also referred to as *plug-and-play priors*, see e.g. [9], [10]), using the parameterization of convolutional neural networks as a regularizer [11], or optimizing over the latent space of a generative model trained on representing the desired type of solutions, see e.g. [12], [13]. Interestingly, such approaches have not yet been exploited for LF reconstruction problems arising from (1), most likely due to the high complexity of light field data.

In this paper, we introduce for the first time, a generative model for light field data for generic reconstruction. The key idea is to model the distribution of light fields using a class of generative autoencoders [14]. Once the training is complete, we use our generative model as a prior in different light field reconstruction problems in an energy minimization framework. Due to the high complexity and variability of the light field data, generating light fields in a consistent fashion is highly challenging. In this paper, we consider only

- *The authors are with with the Department of Computer Science, University of Siegen, Siegen 57076. * indicates equal contribution*



(a) (b) (c)

(d)

Fig. 1. (a) A full 5×5 LF, with central view marked in red. (b) Central view (CV) extracted from (a), with a small patch of this central view marked in blue. (c) This patch passes through a convolutional feature extractor to output central view features (CVF). (d) The encoder E_1 of the CVAE maps an LF patch to a latent variable z, while generator G_1 of the CVAE maps z back to the LF patch using CVF as an additional input.

small baseline light fields and we address this challenge by training generative model for light field patches instead of entire light fields. The advantage of our approach is that the model learned on patches can readily generalize to a variety of scene classes, while being small enough to be amenable for training.

We propose to learn the representation of light field patches with a variational autoencoder conditioned on the central view (CVAE). Fig. 1(d) shows the schematic of the CVAE. The CVAE, consists of an encoder E_1 that takes an LF patch as input and returns a low-dimensional latent code z. The generator G_1 maps this latent code back to the LF patch. A convolutional feature extractor Fig. 1(c) provides features of the central view of the light field patch as an additional input to both the encoder and generator of the CVAE. Consequently, both the encoder and the generator utilize the information from the central patch. In the reconstruction of the light field patch shown in Fig. 1 (d), we observe that the generator can map the encoded latent variable along with the features of the central view to a light field patch which looks similar to the input patch. This indicates that the encoder has learned to encode properties such as disparity and occlusion in the latent space, such that the generator can reconstruct the LF patch just from this latent code and the central view features.

We solve different LF reconstruction problems using our generative model namely, view synthesis, spatial angular super resolution and coded aperture to demonstrate the flexibility of our approach. We illustrate the efficacy of the CVAE in different LF reconstruction tasks when the central view is given. Even when the central view is unavailable, we can exploit the CVAE to aid LF reconstruction. Experimental results indicate that our approach performs close to end-to-end trained networks trained for a specific LF reconstruction tasks, while retaining the flexibility to address different reconstruction tasks. Moreover, our approach can effectively handle different distortions and noise in inputs while learning-based approaches cannot handle such variations without retraining.

2 Related Work

Light field reconstruction

Light field reconstruction has been performed from different observation models, i.e., different instances of (1), such as coded aperture [15], [16], [17], compressed sensing [2], [18], novel view synthesis and angular super-resolution [3], [19], [20], [21], [60], spatial angular super-resolution aided by high resolution central view [22] and also light-field image in-painting and focal stack reconstruction in [23]. Since virtually all such observation models make the solution of (1) an *ill-posed* problem, a natural strategy is to consider regularized energy minimization methods, for example [2], [21]. Alternately, one could estimate depth maps [24], [25] or disparity maps which could be subsequently used to synthesize light fields, see [3], [26] for examples. Recently learning-based approaches have also been applied in LF recovery for coded aperture in [6], [8], compressed sensing in [5], view synthesis and angular super-resolution in [4], [7], [27], [28], [29], [30], spatial and angular super-resolution in [31], [32] as well as view extrapolation for wide baseline light fields in [33], [34].

While neural network-based reconstruction schemes [4], [5], [6], [7], [8], [30], [32], [35] outperform traditional approaches to LF reconstruction by a large margin, they are applicable to specific observation models only, i.e., they are not flexible in adapting to modifications of the observation model. We note that [36] is a deep network-based approach for compressive LF recovery, which also takes a mask as an input to the deep network, achieving flexibility with respect to different masks for compressive sensing.

Learning light field representations has been addressed previously since the data is high dimensional and contains redundant information. Representations based on sparse coding have been utilized to perform inference tasks such as disparity estimation [37], [38] and LF reconstruction [2]. Alperovich *et al.* [39] have shown that an autoencoder trained on stacks of epipolar-plane images (EPI) can learn useful LF representations which can be used for supervised training for disparity estimation and intrinsic decomposition. Recently, there have been efforts to synthesize a light field from a single image in [40], [41], [42]. Srinivasan *et al.* [40] train an end-to-end network which is based on depth estimation from single image and subsequent warping to render light field. CNN-based appearance flow estimation is used in [41], to accomplish LF synthesis from a single image. Chen *et al.* [42] synthesize a light field from single image without estimating any depth map using deep neural network employing GAN loss. Generating a light field from a single view can have several possible solutions. The approaches [40], [41], [42] output a fixed light field for a given input image. In contrast, our CVAE can generate different LF patches for the same input patch, by sampling in the latent distribution.

Generative models

Deep generative models starting from variational autoencoders [43], and GANs [44] have emerged as an important tool for learning data representations in an unsupervised way. These models have demonstrated an impressive ability in generating realistic new image samples from specific image classes [45]. However, training generative models which can synthesize class independent natural images remains difficult and often requires huge network architectures like [46]. Recently, generative models have also been proposed for videos [47], [48]. However, deep generative modeling to capture light field distribution has not yet been attempted.

Image reconstruction using generative models

In addition to generating realistic samples of images [45], [49], generative models have also been used as priors in various image reconstruction [12], [13], [50], and image manipulation [51] tasks. Some of these algorithms involve an optimization in the latent space of the generative model with gradient descent based updates in [12], [13]. More sophisticated optimization schemes such as projected gradient descent, ADMM have also been used in conjunction with GAN priors for optimization in the latent space [52], [53], [54]. Alternatively, encoder-decoder based optimization has also been used with gradient-based updates in [50] and with ADMM in [55]. Such methods have, however, not been exploited for LF data yet.

3 Light Field measurement model

Continuous light fields are represented using the plenoptic function $L(\mathbf{x}, \mathbf{v})$ that denotes the radiance of the scene emitted at the spatial position $\mathbf{x}$ and in the angular direction $\mathbf{v}$. For the discrete light field, we consider the angular resolution for each axis to be N_v, and the spatial resolution of each view to be $N_x \times N_x$. The discrete light field can be represented in vector form as $\mathbf{l} \in \mathbb{R}^k$ with $k = N_x^2 \cdot N_v^2$. In this work, we attempt to solve 3 different LF reconstruction problems utilizing generative priors: (i) LF view synthesis/ view upsampling, (ii) Spatial-angular super-resolution aided by a central view, and (iii) LF recovery from coded aperture images. Among these 3 models, for LF view synthesis and spatial angular super-resolution, we assume that the central view is available. We now consider the specific measurement models for each of these reconstruction tasks.

View synthesis / Angular super-resolution

The task of view synthesis is to recover all sub-aperture images (SAIs) from a sparse subset of input views. The forward model can be considered to be a point-wise multiplication of the light field with a binary mask M, whose value is 1 at the known views, and 0 at all other locations, leading to

$$i(\mathbf{x}, \mathbf{v}) = L(\mathbf{x}, \mathbf{v}) \odot M(\mathbf{x}, \mathbf{v}). \tag{2}$$

where $\odot$ is the point-wise multiplication operator.

Spatial and angular super-resolution using central view

Here the task is to recover all SAIs from a sparse subset of spatially down-sampled input views. Furthermore, we assume that the central view is available in full resolution which aids in spatial upsampling of novel views. The corresponding measurement model can be written as

$$i(\mathbf{x}, \mathbf{v}) = (L(\mathbf{x}, \mathbf{v}) \odot M(\mathbf{x}, \mathbf{v}))_{\downarrow_{s(\mathbf{v})}}. \tag{3}$$

where M is a binary mask which is non-zero only at known views, and $\downarrow_{s(\mathbf{v})}$ is the spatial down-sampling operation of the known views. However, the central view is available at full resolution, i.e the downsampling factor is 1, for the central view.

Coded aperture

Coded aperture images are the result of optical multiplexing only along angular dimension. In a continuous setting, the coded aperture image formation model can be written as

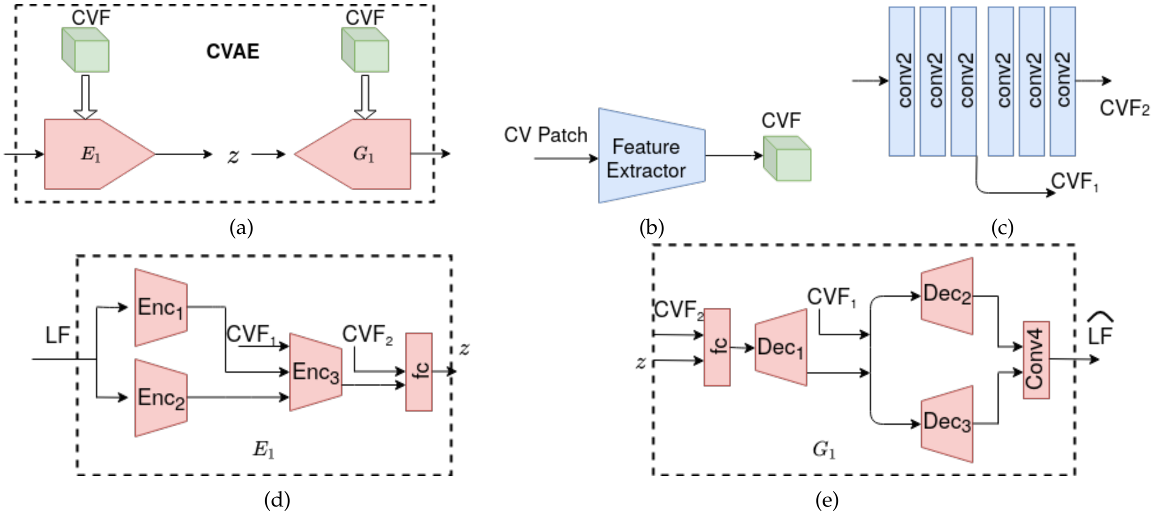
$$i(\mathbf{x}) = \int L(\mathbf{x}, \mathbf{v}) M(\mathbf{v}) d\mathbf{v} \tag{4}$$

where M represents the coded mask, which depends on the angles $\mathbf{v}$, but not on the spatial position.

Each of the forward models given in Eqs. (2), (3), (4),is a linear measurement model, which can be discretized and represented via (1). In the following, we develop a generative model for light fields, which can be exploited for solving such general LF reconstruction problems.

4 Light Field Generative Model

Though light field data has high dimensionality, patches of light fields lie in a manifold of much lower dimension owing to their redundant structure [39]. Therefore, training generative models for LF patches instead of full light fields is a promising alternative. Moreover, the representation learned on the small LF patches can generalize to a wide variety of



(a) (b) (c)

(d) (e)

Fig. 2. (a) Schematic of CVAE. (b) Central view feature (CVF) extraction. (c) Architecture of feature extractor, CVF={CVF$_1$,CVF$_2$}. (d) Schematic of encoder E_1 of CVAE. (e) Schematic of generator G_1 of CVAE

different light fields independent of any specific class of objects.

We introduce generative models for 4D light field patches based on a class of variational autoencoders known as Wasserstein autoencoders [14]. In addition to the autoencoder MSE loss between input and output, these models have a maximum mean discrepency (MMD) penalty between the encoder distribution, and the prior latent distribution, instead of the Kullback-Leibler (KL) divergence penalty found in the traditional variational autoencoders. The loss function is given as

$$\text{Total loss} \; = \; \text{MSE loss} \; + \; \lambda \cdot \text{MMD loss} \qquad (5)$$

We propose a generative model for LF patches, a conditional variational autoencoder (CVAE), conditioned on the central view. We trained the model for LF patches of spatial resolution 25×25. The angular resolution of the LF patch is chosen to be the same as the angular resolution of the light field to be reconstructed (5×5 and 7×7 in our experiments).

4.1 Conditional Generative Model

Although we restrict the spatial extent of a LF patch to 25×25 pixels, due to diverse possibilities of texture content, parallax effects and occlusion effects, representing any patch with a generative model would still be a difficult task. Therefore, we develop a model which is conditioned on the patch corresponding to the central view. With the central patch being fed into the network as an additional input, the encoder only needs to encode the additional information to represent the parallax and occlusion effects in the light field. The decoder learns to utilize the information from the central view to map the latent variable to the light field.

The schematic of the CVAE with its main components is illustrated in Fig. 2. Features of central view are extracted from a convolutional feature extractor at different layers (CVF$_1$ and CVF$_2$), which are together referred to here as the central view features (CVF). These are simultaneously fed to both encoder and generator. The feature extractor is jointly trained along with the encoder and generator. We employ 3D and 2D convolutions in our architecture as an alternative to computationally expensive 4D convolutions. To realize this, the encoder blocks Enc$_1$ and Enc$_2$ in E_1 (Fig. 2 (d)) take the input 4D LF patch as a set of 3D LF patches by splitting them along the horizontal and vertical view dimensions, respectively. The outputs of these encoder blocks are together fed into a common encoder Enc$_3$, along with a set of central view features CVF$_1$. The output of Enc$_3$ together with central view features CVF$_2$ are further encoded by fully connected layers to output latent code z. The generator G_1, takes in the latent code and central view features CVF$_2$ which first pass through linear fully connected layers, followed by a common partial decoder Dec$_1$. This decoder's output together with central view features CVF$_1$, simultaneously pass through the row and column decoders Dec$_2$ and Dec$_3$. These features are together input to a final 4D convolutional layer. Further details of CVAE network architecture for both the conditional models are provided in the supplementary material.

4.2 Reconstruction from Generative Model

To illustrate the performance of the CVAE, Fig. 3 depicts sample reconstructions (encoding and decoding) from our CVAE for 4 LF patches. We handle colored light field inputs by reconstructing each color channel separately. In the second row of Fig. 3, we observe that our CVAE can reconstruct the input LF patches quite accurately. It captures the disparity across different views, and is able to realistically estimate pixel values that are not present in the central view due to the parallax. To demonstrate the efficacy of the CVAE latent code in encapsulating different properties of the input LF patch, we show the generation of a light field from an arbitrarily chosen central patch in the third row of Fig. 3. The latent representation of the LF patch shown in the first row is used for generating this output. As we can see, the result is a new LF patch with disparity values similar to the input LF patch in the first row of Fig. 3. This indicates that the latent vector indeed encodes an understanding of the geometry of the scene. In the following, we develop LF recovery techniques which exploit the strength of our CVAE.

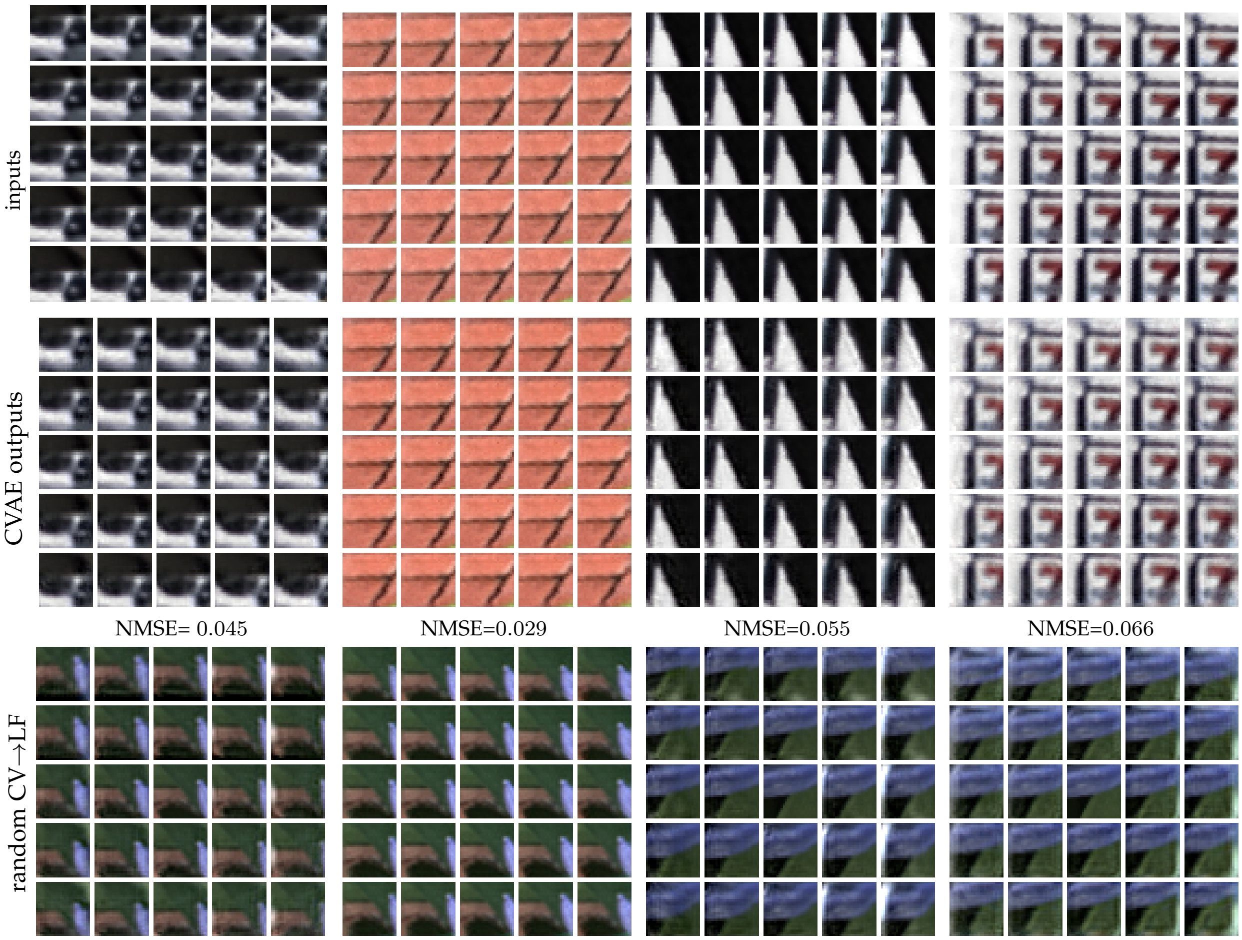


Fig. 3. Sample reconstruction from CVAE. The first two rows are input LF patches and corresponding reconstructions from CVAE. The third row shows the CVAE mapping of an arbitrary central patch to an LF patch with disparity similar to input LF patch, using the latent code corresponding to the second row. Reported numbers are normalized RMSE (NMSE) values of the reconstructions with respect to the corresponding input patches.

5 Generic Light Field Recovery

Light field recovery from measurements as seen in Sec. 3 is an inherently ill-posed problem, and needs strong priors to obtain acceptable solutions. We consider two scenarios: i) the central view is available, and ii) the central view is not available. We now proceed to solve the LF reconstruction problems in both the cases using our CVAE from Sec. 4.

Central view available

In some LF recovery applications such as view synthesis, or spatial angular super-resolution, one can assume that the central view is known. For such scenarios, we utilize our CVAE model for reconstruction. Given the central view, the generator of CVAE is trained to always map a latent code to a light field patch. Therefore, we optimize over the latent space similar to [12], [13] to obtain a latent code that best captures the scene geometry corresponding to the observations. However, unlike [12], [13], we use a conditioned generative model,which additionally takes the central view as input. More specifically, we solve

$$\min_{z} \|\mathbf{i} - \mathbf{\Phi} G_1(z)\|_2^2 \tag{6}$$

where G_1 is the generator of CVAE and $\mathbf{\Phi}$ is the operator corresponding to measurement from angular subsampled views or from spatial and angular subsampled views, assuming the central view is present. We minimize (6) locally using Adam [56], a gradient-based optimization algorithm. After finding a local minimum $\hat{z}$ of (6), $G_1(\hat{z})$ is considered to be our final light field estimate.

Central view not available

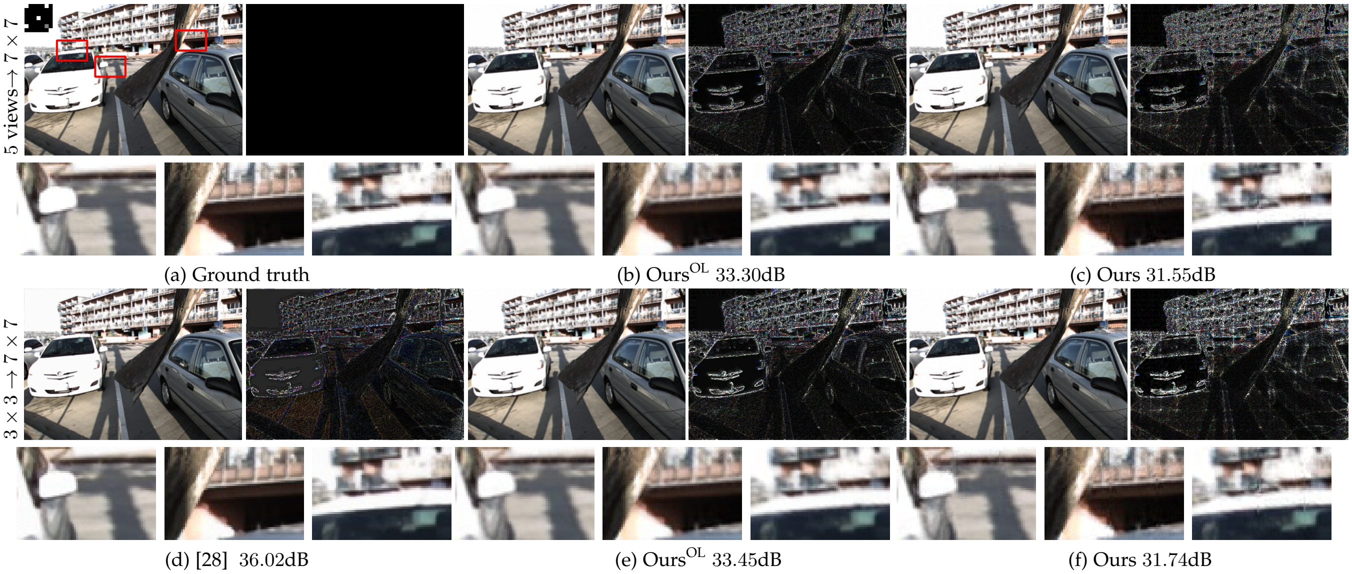
In LF recovery applications such as recovery from coded aperture, the central view is not available. Even in this case, we can utilize the generator of CVAE for reconstruction. The only difference is that we now optimize both for the latent code z and the central view $\mathbf{c}$. We solve the following optimization problem

$$\min_{z,\mathbf{c}} \|\mathbf{i} - \mathbf{\Phi} G_1(z, \mathbf{c})\|_2^2, \tag{7}$$

where $\mathbf{\Phi}$ is the forward measurement operator. We solve this problem using Adam optimizer to obtain local minimizers $\hat{z}$ and $\hat{\mathbf{c}}$. We find our final LF estimate as $G_1(\hat{z}, \hat{\mathbf{c}})$.

6 Experiments

To be able to compare with recent network-based approaches on small baseline light fields, we evaluate view synthesis from sparsely sampled views for LFs with angular resolution 7 × 7. We evaluate LF recovery for view synthesis,



(a) Ground truth (b) Ours^{OL} 33.30dB (c) Ours 31.55dB

(d) [28] 36.02dB (e) Ours^{OL} 33.45dB (f) Ours 31.74dB

Fig. 4. Result of 7×7 view synthesis for the LF 'Cars'. Shown is the novel view at angular location (6,6), depicted as gray location in the inset. The mask for selecting 5 input views is shown in the inset of ground truth view. Figures in the first row a)$-$c) depict ground truth view, and the results of our approach using 5 input views with and without overlapping patches in that order. Figures d)$-$f) in the second row provide visual comparison of novel views generated using approach of Wu *et al.* [28], and our approach using 3×3 angular views. Error maps and zoomed in patches are depicted along with corresponding novel views, with error magnified by a factor of 10. Results best viewed when zoomed in.

spatial-angular super-resolution and coded aperture for LFs with angular resolution 5×5. We will make our code publicly available at https://github.com/KVGandikota/Generative-Light-Field-Models/.

6.1 Experimental Setup

Baselines:

We obtain the performance references for the reconstruction tasks using both, model- and network-based approaches for comparisons. For 7×7 view synthesis, we compare with the recent neural network-based technique of [28]. For comparison with a traditional approach, we report the performance of the depth-based approach from [28].

The dictionary-based approach of Marwah *et al.* [2], developed for compressed sensing, is a flexible technique, which can be used with any observation model. We use their open sourced code[1] which is available for LFs of angular resolution 5×5. We use this as a reference for model-based approaches on all the 3 recovery tasks for 5×5 LFs. For the best performance of [2], we always compute their result obtained by averaging over overlapping patches with stride 1. Additionally, for comparison with a recent neural network baseline, we compare with [60] for 5×5 view synthesis. We use their publicly available code to retrain their model for this task. For reconstruction from coded aperture we compare to the neural network based approach of [6].

Datasets:

For training the generative models, we used the following datasets: i) The training set used by Kalantari *et al.* [4], ii) the training set used in CNN-based depth estimation for light fields by Heber *et al.* [57], and iii) the training set used in encoder-decoder-based light field intrinsics [39]. These datasets contain a significant number of samples with effects such as occlusions and specular reflections. We create a training set by randomly cropping $250K$ LF patches of resolution $5 \times 5 \times 25 \times 25$ in gray scale from these datasets and use them for training the CVAE with angular resolution 5×5. Similarly, a training set of $250K$ LF patches of resolution $7 \times 7 \times 25 \times 25$ was created to train the CVAE with angular resolution 7×7. The datasets from [39] and [57] have high disparity, therefore we down-scale those light fields spatially by a factor of 1.4 before extracting patches from this data. We investigate the effect of training with these datasets by training a separate CVAE on each of them. The comparison of sample reconstructions using these models with our model trained on all the three datasets is provided in the supplementary material. Furthermore, we also study the performance of our generative model for LF patches of different spatial extents, which is provided in the supplementary material.

We evaluate the light field recovery on synthetic and real datasets. Specifically, for LFs of angular resolution 5×5, we evaluate the recovery from all the tasks on the light fields "Dino", "Kitchen", "Medieval 2" and "Tower" from the synthetic New HCI dataset [58]. Furthermore, we evaluate coded aperture reconstruction on the real light field from [6]. We evaluate view synthesis for LFs of angular resolution 7×7 on the test set of [4] which contains 30 real light fields captured by a Lytro Illum. Further, we also evaluate 7×7 view synthesis on the LFs 'Reflective 9', 'Reflective 13', 'Reflective 22', 'Reflective 27', 'Reflective 29', 'Occlusions 16', and 'Bikes12' from Stanford Lytro light field archive [59], which contain significant reflections, transparencies, specularities and occlusions.

[1] http://web.media.mit.edu/~gordonw/CompressiveLightFieldPhotography/

Dataset	$3 \times 3 \rightarrow 7 \times 7$				5 views$\rightarrow 7 \times 7$	
	[28]	Ours	Ours$^{\mathrm{OL}}$	[24]†	Ours	Ours$^{\mathrm{OL}}$
30 scenes [4]	41.16	38.53	39.77	34.42	38.29	39.57
7scenes [59]	41.24	39.62	40.48	-	39.07	40.00

TABLE 1
Average PSNR of novel views in dB for 7×7 view synthesis. † indicates PSNR values of [24] are as reported in [28].

	Clean	$\sigma = 0.05$	$\sigma = 0.1$	S&P	50% pixels
[28]	36.02	33.34	29.95	25.02	13.60
Ours	31.74	31.75	31.67	31.66	31.68
Ours$^{\mathrm{OL}}$	33.45	33.47	33.41	33.35	33.39

TABLE 2
$3 \times 3 \rightarrow 7 \times 7$ view synthesis result on the LF 'Cars', when input views other than central view are corrupted. Shown are PSNR values in dB

Generative model training:

We used Pytorch 1.1.0 for all our experiments. For training the CVAEs, we use mini-batches of size 128 and trained the models for 5×5 and 7×7 views with spatial extent of 25×25 pixels for 150 epochs. We used Adam optimizer [56], with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We set the initial learning rate to 10^{-3}, which is decreased by a factor of 2 after 30 epochs, further by a factor of 5 after first 50 epochs and finally by a factor of 10 after 100 epochs. For both the models, we choose the factor λ in eq. (5) to be 100.

LF recovery:

Since our generative models are trained on gray scale patches, we divide the input into patches of suitable dimensions and use our generative models on all color channels separately. We initialize the latent code z with a random sample drawn from the same posterior distribution that was used for the latent space during the training of the generative model (i.e. isotropic Gaussian with variance of 2). We observed that different random initializations of z lead to similar quality of reconstruction. For recovery from coded aperture, the central view is not available. In this case, we initialize the central view with the coded image itself scaled between 0 and 1. We solve the LF reconstruction tasks using Adam optimizer as discussed in Sec. 5, until convergence.

6.2 Results

We now evaluate the efficacy of our approach on different LF recovery tasks. We perform quantitative evaluation in terms of PSNR and also qualitative evaluation by comparing light field views of our approach with ground truth and baseline methods and show the corresponding error maps. Additional visual comparisons and videos of the reconstructed LFs are provided in the supplementary material.

6.2.1 Central View Available

View synthesis 7×7*:*

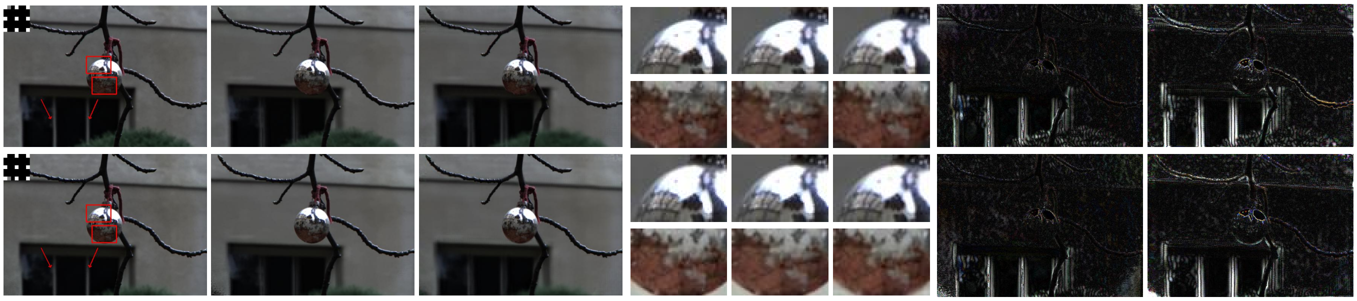
We compare our approach with recent CNN-based technique of Wu *et al.* [28] for LF reconstruction from sparsely sampled input views. We consider upsampling the angular resolution from 3×3 to 7×7. Since central view is available for this task, our approach uses CVAE for reconstruction. We use the publicly available trained model of [28][2] for evaluating their approach. We also report the performance of a traditional depth estimation-based approach from [28] for this task, where the depth is estimated using the approach of Jeon *et al.* [24], followed by a novel view synthesis by warping the input views following [26]. Apart from the specific case of 3×3 input views, our method can still be applicable if any arbitrary set of views are given as input along with the central view. To demonstrate this flexibility, we also show 7×7 LF reconstruction from 5 randomly chosen input views including the central view. The mask used for selecting the 5 input views is provided in the inset of Fig. 4 a). Since view extrapolations cannot be handled by Wu *et al.* [28], we show visual comparison only with the ground truth for this task.

Results of our quantitative evaluation on 30 real LFs of Kalantari [4] test set and 7 scenes selected from Stanford Lytro dataset [59] are provided in Tab. 1. 'Ours$^{\mathrm{OL}}$' indicates our reconstruction using overlapping patches with stride 5. Following Wu *et a.l* [28], we show the result of average PSNR of the luminance component of novel synthesized views. For brevity, we report only average PSNRs of the LFs in each test set. Quantitative comparisons for individual LFs are provided in the supplementary material. For the task of view upsampling from 3×3 to 7×7, we compute the average PSNRs of the 40 novel views. For this task, we find that our performance is approaching the CNN-based method of [28], with a PSNR reduction of only 1.4 dB when we use overlapping patches, and 2.6 dB when non-overlapping patches are used on Kalantari test set [4]. Our approach also outperforms the depth-based approach using the method of Jeon *et al.* [24] by a large margin. Further, our performance is close to the method of [28] on the scenes selected from [59] as well, with PSNR reduction of only 0.8 dB and 1.6 dB respectively, when overlapping and non-overlapping patches are used. Even when the number of known views is reduced to 5, our average PSNR of 44 novel views is 39.57 dB on the 30 scenes [4] with a reduction of only 0.2 dB, and average PSNR of 40.00 dB with a reduction of 0.48 dB on the 7 scenes from [59], demonstrating the strength of our approach.

A qualitative comparison of the synthesized views for the task of 7×7 view synthesis is provided in Fig. 4 for the LF 'Cars' from the 30 scenes test set. The newly synthesized view at angular location $(6, 6)$ (depicted by gray location in the inset) are shown. The first row of Fig. 4 (a)$-$(c) gives a visual comparison of the results of our approach with the ground truth when 5 input views are used. Visually, it can be seen that our approach provides a reasonable reconstruction quality even when using a limited number of input views. The second row of Fig. 4 (d)$-$(f) compares our method with the approach of Wu *et al.* [28], for the task of $3 \times 3 \rightarrow 7 \times 7$ angular super resolution. In terms of reconstruction quality, our approach performs slightly worse than [28]. However, this is to be expected as [28] uses network specifically trained for this task. In contrast, we obtain a comparable reconstruction quality with flexible input views. It can be noticed from the error maps and zoomed in patches that our approach preserves the details fairly well. Further, we can observe that there are errors at the patch boundaries when non-overlapping are used. These errors are reduced due to averaging effect when overlapping patches are used.

In Fig. 5, we illustrate our reconstruction of the LF

[2] https://github.com/GaochangWu/lfepicnn



a) Ground truth b) [28] 35.76 dB c) Ours$^{\text{OL}}$ 35.06 dB a) b) c) b) [28] c) Ours$^{\text{OL}}$

Fig. 5. Visual comparison of our synthesized views (depicted by gray locations in the inset) for the task of $3 \times 3 \rightarrow 7 \times 7$ view synthesis for the LF 'Reflective13'. Columns $1-3$ depict a) the ground truth views, the result using b) Wu *et al.* [28] and c) our approach using overlapping patches respectively. Columns $4-6$, the patches of columns $1-3$. Columns $7-8$ depict the error maps corresponding to columns $2-3$ with error magnified by a factor of 10. The brightness of the zoomed in patches is increased for better illustration. Average PSNR in dB of 40 novel views is shown.

'Reflective13' from the Stanford Lytro dataset and compare it with the approach of [28] for the task of $3 \times 3 \rightarrow 7 \times 7$ view synthesis. The novel synthesized views at angular locations $(1, 2)$ and $(7, 2)$ (depicted by gray location in the inset of ground truth) are shown. This is a challenging scene which contains a highly reflective ball in the foreground, and high disparities (4 pixels between adjacent views) in the background. We can observe from the synthesized views and corresponding error maps that our approach provides reconstructions which are slightly worse than [28]. On closer inspection of the zoomed in patches, we can observe that our method can reconstruct well the reflections which are slowly varying across views (patches on the top in each row). We observe reasonable reconstruction even when there is a high variability in the reflections across views (patches on the bottom). However, our approach cannot handle high disparities in the background causing ghosting artifacts, as seen in the corresponding regions in the reconstructions, which are indicated by red arrows in the ground truth views. We observe that the approach [28] also cannot handle such large disparities, which are also evident in the error maps.

To further demonstrate our flexibility vis-a-vis end to end trained networks, we consider the task of $3 \times 3 \rightarrow 7 \times 7$ angular super resolution and compare our reconstruction with Wu *et al.* [28], when inputs are corrupted. We assume that the central view is clean and the remaining 8 views are corrupted by different distortions. The qualitative and quantitative comparison of our reconstructions with the approach of Wu *et al.* [28], with corrupted input views is provided in Fig. 6 and in Tab. 2 for the LF 'Cars'. The reconstructed view at angular location $(6, 6)$ is depicted. With additive Gaussian noise of standard deviation $\sigma = 0.05$ in 8 input views, the PSNR of the reconstructed views using [28] drops from 36.02 dB to 33.34 dB. When we increase the noise level to $\sigma = 0.1$ this value further drops to 29.95 dB. This degradation in the quality of reconstruction is also evident from the error maps in Fig. 6. In contrast, our reconstruction quality is robust to addition of noise.

We also consider corruption of input views with salt-and-pepper noise with a probability of 0.05. Even in this case, the performance of [28] is severely affected, with PSNR reduction of 11 dB compared to the clean case, where as our performance only shows a marginal decrease of 0.1 dB.

LF	Mask M_1				Mask M_2			
	Ours	Ours$^{\text{OL}}$	[2]	[60]	Ours	Ours$^{\text{OL}}$	[2]	[60]
Dino	39.57	41.53	34.61	43.68	38.18	39.83	32.99	42.46
Kitchen	33.59	34.95	30.80	37.01	33.06	34.41	29.83	36.29
Medieval2	34.86	35.94	32.19	36.75	34.55	35.66	31.51	36.25
Tower	31.24	32.30	28.45	34.00	30.28	31.31	27.67	32.97

TABLE 3
5×5 View Synthesis: PSNR values in dB

We note that we employ an L_1 loss, as it is more suited to handle salt and pepper noise when compared to the traditional L_2 loss in Eq.(6). This demonstrates the flexibility of our energy minimization-based approach in adapting to different noise statistics. When we use an L_2 loss instead, our PSNR dropped by about 2 dB compared to the clean case. Finally, when 50% pixels are randomly dropped from the 8 known views, the neural network-based approach of [28], completely fails in reconstruction. In contrast, we can incorporate an additional mask corresponding to the missing pixel locations in our optimization, and consequently our reconstructions remain robust to this distortion. We can also accomplish LF recovery when the input views, including the central view are corrupted. As shown in Eq. (7), this requires optimizing jointly for the latent code and the central view. We demonstrate with additional experiments in the supplementary material that our approach can provide a reasonable reconstruction, even when the central view is significantly corrupted. We find that using an additional total-variation (TV) penalty on the central view further improves our performance under noise.

View synthesis 5×5*:*

We compare our approach for view synthesis with [2] and [60] for two different input views using masks M_1 and M_2. For evaluating the performance of [60] we use separate networks trained end-to-end for view synthesis with each of the masks. A qualitative comparison of the synthesized views is provided for the LF 'Dino' for mask M_1 and M_2 in Fig. 7. The locations of known views are depicted in white in the inset of Fig. 7, and gray represents the location of the reconstructed view. Extrapolating novel views away from known views is difficult. Even for this challenging case, we observe the quality of our reconstruction with both, overlapping and non-overlapping patches, is better and

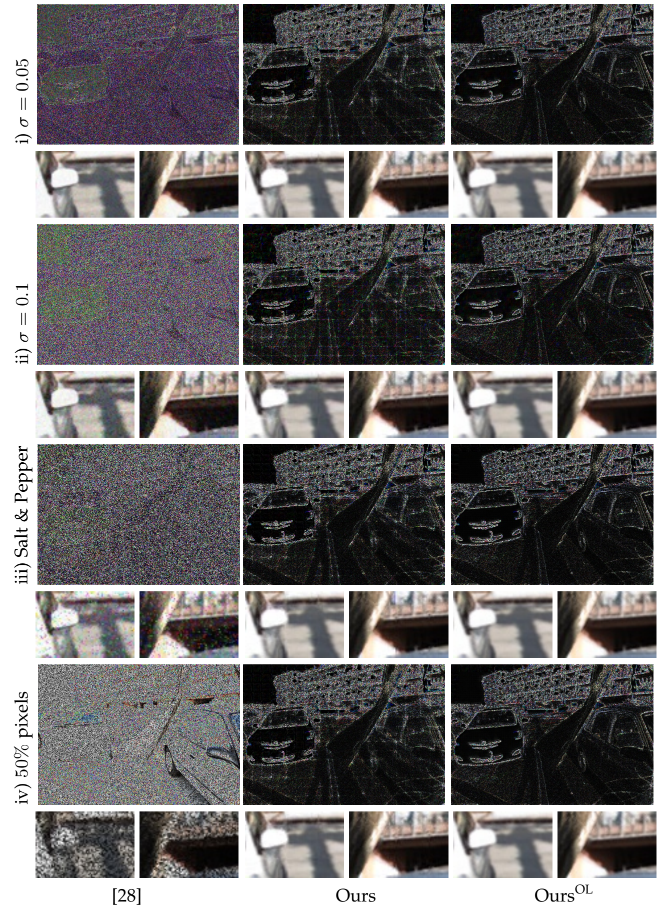


Fig. 6. Novel view at angular location $(6,6)$ for the task $3 \times 3 \to 7 \times 7$ view synthesis. Columns $1-3$ depict the result using Wu *et al.* [28] and our approach using non-overlapping patches and overlapping patches respectively. Shown are the zoomed in patches of the reconstructed views and error maps with error magnified by a factor of 10. Among the 3×3 input views, central view is clean. For the the remaining 8 views, we consider the following corruptions (rows i$-$iv) i) additive Gaussian noise $\sigma = 0.05$. ii) additive Gaussian noise $\sigma = 0.1$ iii) salt and pepper noise with a probability of occurrence of 0.05. iv) 50% pixels randomly dropped from views. Results best viewed by zooming in.

sharper compared to the reconstruction from the dictionary-based approach of [2]. The neural network approach of [60] provides even better reconstruction, which is expected with end-to-end networks specifically trained for each of the masks. This is also evident from the error maps shown in Fig. 7. We can observe that averaging effect of overlapping patches mitigates the errors at the patch boundaries in comparison to our approach without overlapping patches.

The results of our quantitative evaluation on synthetic HCI data are summarized in Tab. 3, where the PSNR of the reconstructed light fields is presented. Our approach without considering overlapping patches is superior by 2.63 dB and 3.13 dB to the dictionary-based approach of [2] with overlapping patches with stride 1, for masks M_1 and M_2, respectively in terms of average PSNR. Our performance further improves when we consider overlapping patches with stride 5, where our approach is better by 4 dB and 4.4 dB, respectively for M_1 and M_2. Further, the neural network based approach of [60] performs the best, with an improvement in average PSNR of 1.69 dB compared to our approach with overlapping patches for both M_1 and M_2.

LF	Mask M_1			Mask M_2		
	Ours	Ours$^{\text{OL}}$	[2]	Ours	Ours$^{\text{OL}}$	[2]
Dino	37.18	39.71	33.07	35.84	38.11	31.70
Kitchen	31.60	33.30	28.98	30.95	32.67	28.10
Medieval2	33.27	34.87	33.26	32.78	34.50	30.26
Tower	29.95	31.15	27.93	28.99	30.23	26.93

TABLE 4
Spatial-angular super-resolution: PSNR values in dB

Spatial and angular super-resolution 5×5*:*

Fig. 8 provides a visual comparison of our LF reconstruction with the approach of [2] for the task of spatial-angular super-resolution on the LF 'Kitchen'. The masks used for the measurements are provided in the inset of ground truth view of the LF 'Kitchen' in Fig. 8. The central view is available in full resolution and is depicted in white. Views in red are spatially down-sampled by a factor of 3. It can be observed that our reconstruction of the novel view (depicted in gray in the inset) with both overlapping patches and non-overlapping patches is of superior quality compared to the reconstruction from the approach of [2]. This is further substantiated by the error maps shown in the Fig. 8, which depict a much lower error in our reconstruction.

Tab. 4 provides a quantitative comparison of our method with the dictionary-based approach of [2]. Again, on average our approach outperforms the approach of Marwah *et al.* [2] by more than 2 dB without, and by more than 4 dB with overlapping patches.

6.2.2 Central View Unavailable

Coded aperture 5×5*:*

We evaluate the LF recovery from 2 coded aperture observations for our approach, [6] and [2], using two different coded mask sets 'Normal' and 'Rotated' (available from [6]), and denote them by M_1 and M_2, respectively. The quantitative evaluation on synthetic data is summarized in Tab. 5. To evaluate the approach of [6], we use the publicly available trained reconstruction network corresponding to M_1. For M_2, we reproduce the values reported in [6], since a trained network is not publicly available. Even without overlapping patches, our method gives superior PSNR values when compared to the model-based approach of [2], with improvement of 1.6 dB for both M_1 and M_2. However, our method is worse by 2.7 dB and 2.3 dB for M_1 and M_2 when compared to [6]. When we use overlapping patches with stride 5, the average PSNR on the test set for our method is comparable to the end-to-end trained model of [6] and is better by 3.97 dB and 3.71 dB for M_1 and M_2 when compared to [2]. For qualitative evaluation, we show sample LF reconstructions using coded masks M_1 on the LFs 'Dino' and 'Medieval' in Fig. 9. We can observe that our approach provides a reasonably good recovery, with performance comparable to an end-to-end trained network. Our recovery is also more accurate when compared to [2].

To demonstrate the vulnerability of the end-to-end trained reconstruction pipeline, we altered the coded aperture mask from the set of M_1 and then perform LF recovery using the method of [6]. Minor changes were applied to only

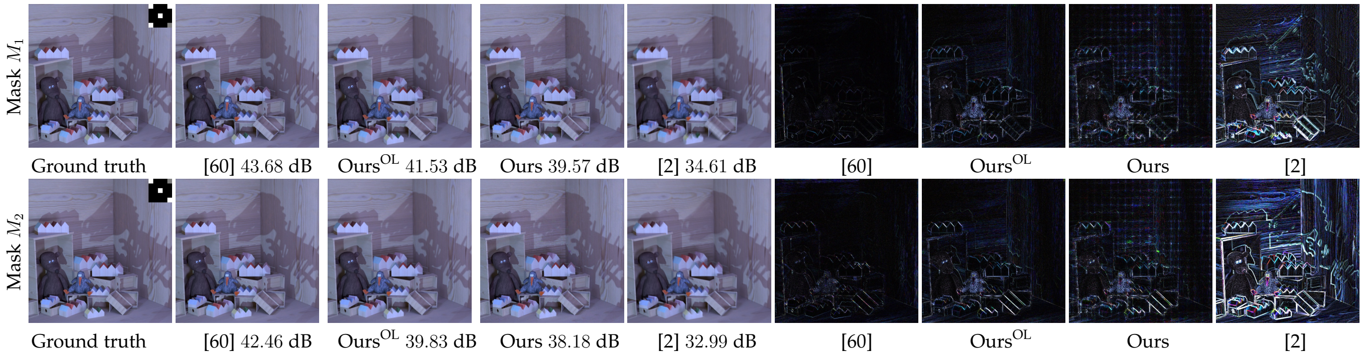


Fig. 7. Result of view synthesis of the LF 'Dino'. Masks M_1 and M_2 are provided as inset of the ground truth views. The columns $1-5$ show the views depicted by gray location in the inset corresponding to i) the ground truth, and synthesized novel views using ii) the method of [60] iii)$-$iv) our approach with overlapping patches and without overlapping patches and v) the method of [2], respectively. Columns $6-9$ illustrate the error maps corresponding to the reconstructed views in columns $2-5$, with errors magnified by a factor of 10. Shown are the PSNR values in dB of the reconstructed LFs. (Results best viewed zoomed in).

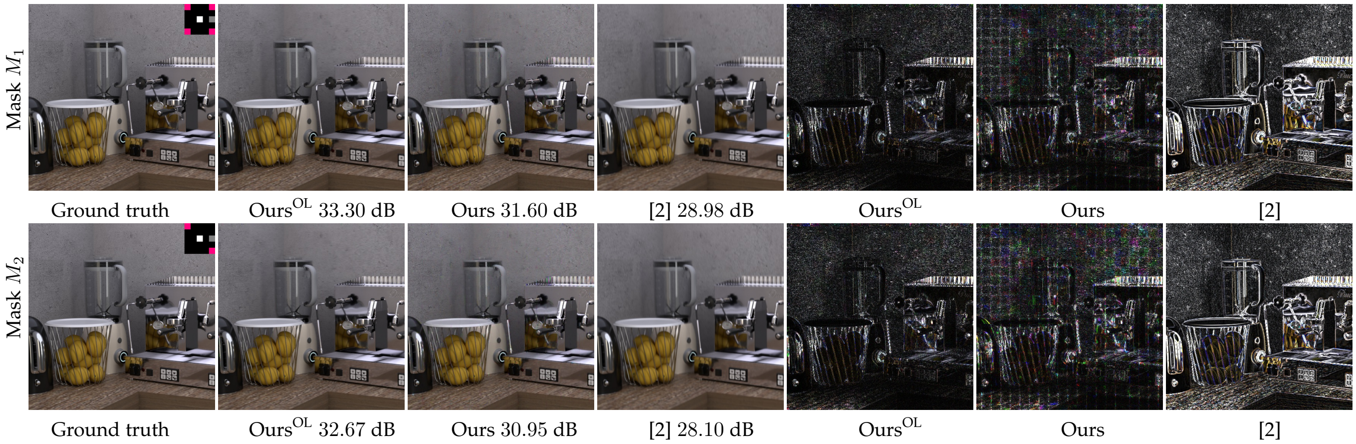


Fig. 8. Result of spatial angular super-resolution of the LF 'Kitchen'. Masks M_1 and M_2 are provided as inset of the ground truth views. Central view in full resolution is depicted in white. Measurements at the locations in red are spatially down-sampled by a factor of 3. The columns $1-4$ from left to right show the views depicted by gray location in the inset corresponding to the i) ground truth, and synthesized views using ii)$-$iii) our approach with overlapping patches and without overlapping patches, and iv) approach of [2]. Columns $5-7$ illustrate the error maps corresponding to the reconstructed views in columns $2-4$, with error magnified by a factor of 10.(Results best viewed zoomed in)

LF	Mask M_1				Mask M_2			
	Ours	Ours$^{\text{OL}}$	[6]	[2]	Ours	Ours$^{\text{OL}}$	[6]†	[2]
Dino	34.97	38.46	38.7	33.28	34.34	38.0	37.5	32.86
Kitchen	31.07	33.29	33.78	29	31.03	33.14	33	29.40
Medieval2	32.90	35.19	34.74	31.37	32.49	34.84	34	31.42
Tower	29.02	30.43	31.63	27.81	28.47	29.86	31	27.33

TABLE 5
Coded aperture reconstruction: PSNR values in dB. [6]† indicates approximate PSNR values for the mask M_2 are taken from [6].

one of the two masks in the set M_1. First, we swap the values of the mask at locations with coordinates $(0,0)$ and $(0,2)$. With this tiny change, the performance of [6] dropped from 38.7 db to 24.3 db on the 'Dino' LF. When we swap the values at three sets of location, the method of [6] completely failed to reconstruct a meaningful light field (yielding a PSNR of 12.2 dB). In contrast, the effect of these changes on our approach is marginal, since our optimization scheme explicitly takes the mask as an input. With the first swap in the mask, our PSNR changed to 38.52 dB, compared to 38.46 dB of the original mask, when we use overlapping patches. With three swaps, the PSNR value for our reconstruction is 38.19 dB, demonstrating our flexibility. Views from the reconstructed LFs are shown in Fig. 10.

We apply our reconstruction method on the real observations obtained in the work of [6]. In their setup, the black-aperture image was not completely dark. Consequently, the image obtained from the black aperture was subtracted from the observations. In Fig. 11, we show a specific view obtained from our reconstruction along with the corresponding result obtained by the authors of [6]. Close-ups near the occlusion boundaries for two different views (with appropriate vertical alignment) in Fig. 11 (c) and (d) show a comparable quality of our approach (left columns) to the results obtains by [6] (right columns).

We also considered other model-based approaches [21], [23] for comparison. We note that these works have not considered view synthesis with arbitrary masks or coded aperture reconstruction. As [21] uses an iterative approach that regularizes the epipolar plane images, it works well with a regular pattern of input views. We found it not to be directly applicable for view extrapolations while our model remains flexible with respect to the pattern of input views. Moreover, we found that [23] crucially depends on a good initial estimate for view extrapolation. Finally, we found that

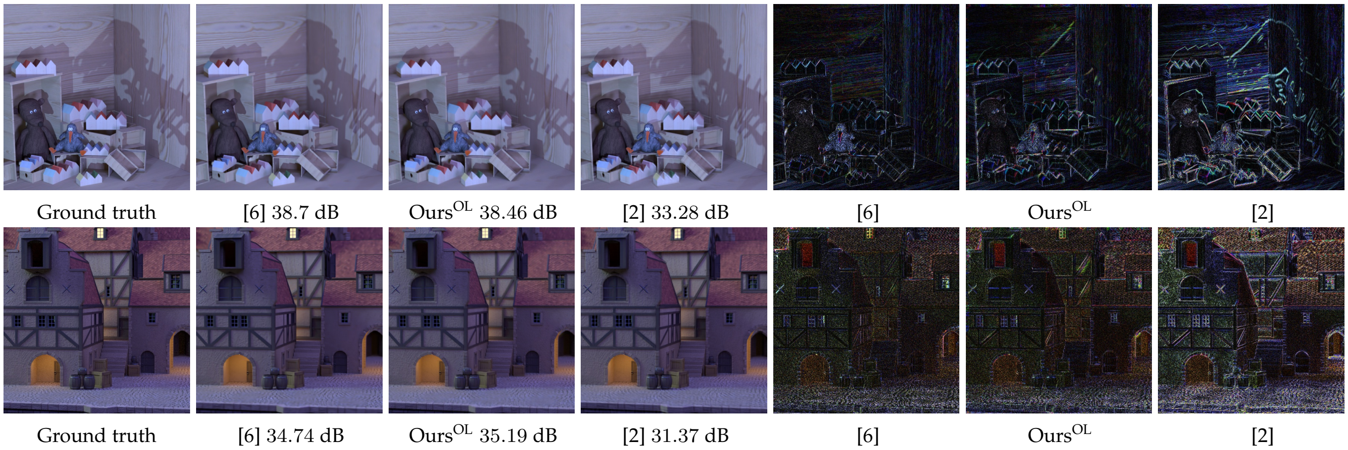


Fig. 9. Coded aperture reconstruction using the coded mask M_1 of [6]. The column 1 depicts the the bottom right ground truth LF view. Columns $2-4$ depict the reconstructed views using [6], our approach and [2] respectively. The error maps corresponding to the views in columns $2-4$ are illustrated in the columns $5-7$, with errors magnified by a factor of 10. PSNR values of recovered LFs are shown. (Results best viewed zoomed in).

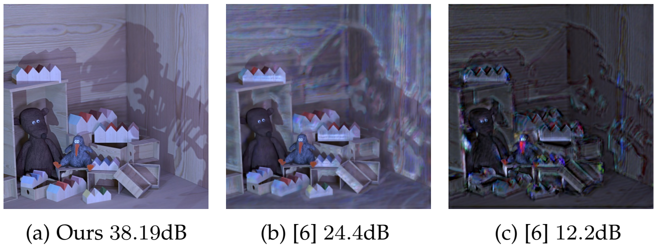


Fig. 10. Effect of minor alterations to the coded mask on reconstruction. Shown is the top left view. (a) Our reconstruction with 3 swaps in the mask. (b) Reconstruction using [6] with 1 swap. (c Reconstruction using [6] with 3 swaps.

the performance of [23] on coded aperture reconstruction was worse than Marwah *et al.* [2]. Therefore, we have not included these comparisons in our results.

Due to patch based processing, and optimization steps required to reconstruct light fields, our reconstruction times are longer. For 7×7 view synthesis, our approach takes nearly 12 minutes on a Nvidia GeForce RTX 2080 Ti machine to reconstruct a full Lytro image (of size $376 \times 541 \times 3 \times 7 \times 7$), which requires 150 update steps per patch. For 5×5 view synthesis and spatial-angular super-resolution, our approach requires 12 minutes to reconstruct LF of size $512 \times 512 \times 3 \times 5 \times 5$, when 250 update steps per patch are used. When overlapping patches with stride 5 are used, our reconstruction times increase by a factor of 25. Another limitation of our approach is that our reconstructions are not satisfactory when the disparity between adjacent views is greater than two pixels. Since the spatial extent of our generative models is only 25×25, it is difficult for our model to capture large disparities in a low-dimensional latent representation, as the views tend to be significantly different. To overcome this limitation, one needs to train a generative model with higher capacity by using LF patches of larger spatial extent. Since our work is the first attempt to develop generative light field models, we consider this to be beyond the scope of this work.

7 Conclusion

We developed the first autoencoder-based generative model conditioned on the central view for 4D light field patches for generic reconstruction. We developed algorithms for generic light field reconstruction by exploiting the strengths of our generative model and evaluated our approach on three different LF reconstruction tasks. Experimental results indicate that our approach leads to high quality reconstructions with a performance superior to other optimization-based approaches, while being only slightly worse but significantly more flexible and robust than end-to-end trained networks. We believe that our experimental results are very promising and can serve as a starting point for further research on generative light field models.

References

[1] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, “Light field photography with a hand-held plenoptic camera,” *Stanford Tech. Report CTSR*, vol. 2005, no. 2, pp. 1–11, 2005.
[2] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, “Compressive light field photography using overcomplete dictionaries and optimized projections,” *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 46, 2013.
[3] S. Wanner and B. Goldluecke, “Variational light field analysis for disparity estimation and super-resolution,” *IEEE transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, no. 3, pp. 606–619, 2013.
[4] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, “Learning-based view synthesis for light field cameras,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, p. 193, 2016.
[5] M. Gupta, A. Jauhari, K. Kulkarni, S. Jayasuriya, A. Molnar, and P. Turaga, “Compressive light field reconstructions using deep learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017, pp. 11–20.
[6] Y. Inagaki, Y. Kobayashi, K. Takahashi, T. Fujii, and H. Nagahara, “Learning to capture light fields through a coded aperture camera,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434.
[7] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 137–152.
[8] A. K. Vadathya, S. Girish, and K. Mitra, “A unified learning based framework for light field reconstruction from coded projections,” *IEEE Transactions on Computational Imaging (TCI)*, 2019.

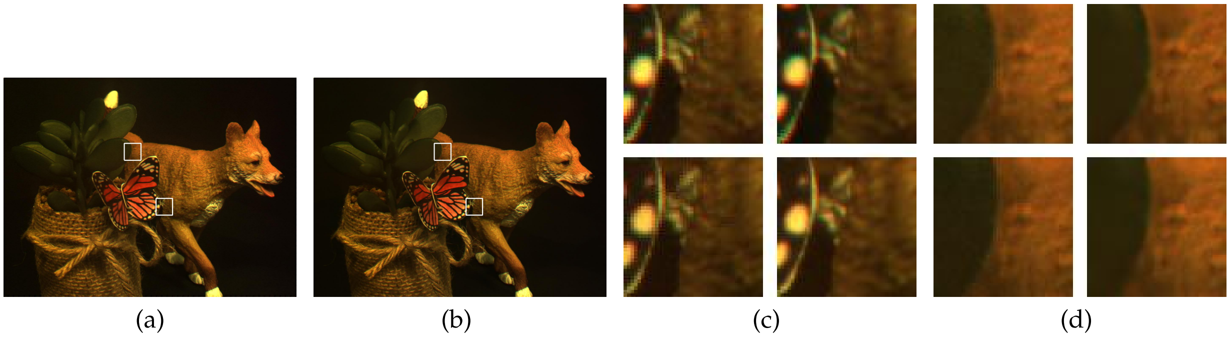


Fig. 11. Real result using the observation of [6]. (a) Central view from our reconstructed light field. (b) Corresponding view from the result of [6]. (c) and (d) left half shows patches from two different views of our reconstruction and right half similarly shows patches from the result of [6].

[9] T. Meinhardt, M. Moller, C. Hazirbas, and D. Cremers, “Learning proximal operators: Using denoising networks for regularizing inverse imaging problems,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1781–1790.

[10] J. Rick Chang, C.-L. Li, B. Poczos, B. Vijaya Kumar, and A. C. Sankaranarayanan, “One network to solve them all–solving linear inverse problems using deep projection models,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[11] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9446–9454.

[12] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, “Compressed sensing using generative models,” in *Proceedings of the 34th International Conference on Machine Learning-Volume (ICML)*. JMLR. org, 2017, pp. 537–546.

[13] Y. Li, S. Liu, J. Yang, and M.-H. Yang, “Generative face completion,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3911–3919.

[14] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf, “Wasserstein auto-encoders,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[15] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen, “Programmable aperture photography: Multiplexed light field acquisition,” *ACM Transactions on Graphics (TOG)*, vol. 27, no. 3, pp. 55:1–55:10, Aug. 2008. [Online]. Available: http://doi.acm.org/10.1145/1360612.1360654

[16] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, “Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing,” *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, p. 69, 2007.

[17] S. D. Babacan, R. Ansorge, M. Luessi, P. R. Mataran, R. Molina, and A. K. Katsaggelos, “Compressive light field sensing,” *IEEE Transactions on Image Processing (TIP)*, vol. 21, no. 12, pp. 4746–4757, 2012.

[18] A. Ashok and M. A. Neifeld, “Compressive light field imaging,” in *Three-Dimensional Imaging, Visualization, and Display 2010 and Display Technologies and Applications for Defense, Security, and Avionics IV*, vol. 7690. International Society for Optics and Photonics, 2010, p. 76900Q.

[19] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, “Light field reconstruction using sparsity in the continuous fourier domain,” *ACM Transactions on Graphics (TOG)*, vol. 34, no. 1, p. 12, 2014.

[20] D. C. Schedl, C. Birklbauer, and O. Bimber, “Directional super-resolution by means of coded sampling and guided upsampling,” in *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2015, pp. 1–10.

[21] S. Vagharshakyan, R. Bregovic, and A. Gotchev, “Light field reconstruction using shearlet transform,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 1, pp. 133–147, 2018.

[22] Y. Wang, Y. Liu, W. Heidrich, and Q. Dai, “The light field attachment: Turning a dslr into a light field camera using a low budget camera ring,” *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, vol. 23, no. 10, pp. 2357–2364, 2016.

[23] C. J. Blocker and J. A. Fessler, “Blind unitary transform learning for inverse problems in light-field imaging,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019.

[24] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1547–1555.

[25] M. S. M. Sajjadi, R. Köhler, B. Schölkopf, and M. Hirsch, “Depth estimation through a generative model of light field synthesis,” in *Proceedings of the 38th German Conference on Pattern Recognition (GCPR)*, ser. Lecture Notes in Computer Science, vol. 9796. Springer International Publishing, Sep. 2016, pp. 426–438.

[26] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis, “Depth synthesis and local warps for plausible image-based navigation,” *ACM Transactions on Graphics (TOG)*, vol. 32, no. 3, Jul. 2013.

[27] G. Wu, Y. Liu, Q. Dai, and T. Chai, “Learning sheared epi structure for light field reconstruction,” *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 7, pp. 3261–3273, July 2019.

[28] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, “Light field reconstruction using convolutional network on epi and extended applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 7, pp. 1681–1694, 2019.

[29] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 333–348.

[30] J. Navarro and N. Sabater, “Learning occlusion-aware view synthesis for light fields,” *ArXiv*, vol. abs/1905.11271, 2019.

[31] 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 (TIP)*, vol. 27, no. 5, pp. 2146–2159, 2018.

[32] N. Meng, H. K. So, X. Sun, and E. Lam, “High-dimensional dense residual convolutional neural network for light field reconstruction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–1, 2019.

[33] P. P. Srinivasan, R. Tucker, J. T. Barron, R. Ramamoorthi, R. Ng, and N. Snavely, “Pushing the boundaries of view extrapolation with multiplane images,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 175–184.

[34] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, “Local light field fusion: Practical view synthesis with prescriptive sampling guidelines,” *ACM Transactions on Graphics (TOG)*, 2019.

[35] N. Meng, T. Zeng, and E. Y. Lam, “Spatial and angular reconstruction of light field based on deep generative networks,” in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019.

[36] O. Nabati, D. Mendlovic, and R. Giryes, “Fast and accurate reconstruction of compressed color light field,” in *Proceedings of the IEEE International Conference on Computational Photography (ICCP)*, 2018, pp. 1–11.

[37] S. Heber and T. Pock, “Shape from light field meets robust pca,” in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 751–767.

[38] O. Johannsen, A. Sulc, and B. Goldluecke, “What sparse light field coding reveals about scene structure,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldluecke, “Light field intrinsics with a deep encoder-decoder network,” in *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9145–9154.

[40] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, “Learning to synthesize a 4d rgbd light field from a single image,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2243–2251.

[41] A. Ivan, I. K. Park *et al.*, “Synthesizing a 4d spatio-angular consistent light field from a single image,” *arXiv preprint arXiv:1903.12364*, 2019.

[42] B. Chen, L. Ruan, and M.-L. Lam, “Lfgan: 4d light field synthesis from a single rgb image,” in*ACM Transactions on Multimedia Computing Communications an Applications*, vol. 16, no. 1, Feb. 2020.

[43] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[44] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proceedings of the Advances in neural information processing systems (NIPS)*, 2014, pp. 2672–2680.

[45] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[46] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[47] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, “Mocogan: Decomposing motion and content for video generation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1526–1535.

[48] A. Clark, J. Donahue, and K. Simonyan, “Adversarial video generation on complex datasets,” in *arXiv preprint arXiv:1907.06571*,2019.

[49] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[50] P. Chandramouli and K. Vaishnavi Gandikota, “Blind single image reflection suppression for face images using deep generative priors,” in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[51] D. Bau, H. Strobelt, W. Peebles, J. Wulff, B. Zhou, J.-Y. Zhu, and A. Torralba, “Semantic photo manipulation with a generative image prior,” *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 59:1–59:11, Jul. 2019.

[52] V. Shah and C. Hegde, “Solving linear inverse problems using gan priors: An algorithm with provable guarantees,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4609–4613.

[53] C. Hegde, “Algorithmic aspects of inverse problems using generative models,” in *the Proceedings of 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2018, pp. 166–172.

[54] F. Latorre, A. Eftekhari, and V. Cevher, “Fast and provable admm for learning with generative priors,” in *Proceedings of the Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 12 004–12 016.

[55] S. Xu, S. Zeng, and J. Romberg, “Fast compressive sensing recovery using generative models with structured latent variables,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2967–2971.

[56] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[57] S. Heber and T. Pock, “Convolutional networks for shape from light field,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3746–3754.

[58] “Heidelberg collaboratory for image processing: 4d light field dataset,” http://hci-lightfield.iwr.uni-heidelberg.de/, 2018.

[59] A. Sunder Raj, M. Lowney, R. Shah, and G. Wetzstein, “Stanford Lytro Light Field Archive,” http://lightfields.stanford.edu/LF2016.html, 2016.

[60] J. Jin, J. Hou, H. Yuan, S. Kwong, “Learning Light Field Angular Super-Resolution via a Geometry-Aware Network,” in *Proceedings of AAAI Conference on Artificial Intelligence*, 2020, pp. 11141-11148.

Paramanand Chandramouli received the Ph.D. degree from Indian Institute of Technology, Madras, India, in 2013. He is currently a Postdoctoral Researcher at the Computer Graphics and Multimedia Systems Group, University of Siegen, Germany. His research interests include computer vision, computational photography, deep learning.

Kanchana Vaishnavi Gandikota received her Masters degree in Electrical Engineering from Indian Institute of Technology, Madras, India, in 2015. She is currently working towards her Ph.D at the Computer Vision Group, University of Siegen. Her research interests include computer vision, image processing and optimization.

Andreas Goerlitz received his Masters degree in Computer Science from Darmstadt University of Technology, Germany, in 2015. He is currently working towards his Ph.D at the Computer Graphics and Multimedia Systems Group, University of Siegen, Germany. His research interests are image processing and computer vision, including light-fields, motion estimation, and 3D reconstruction.

Andreas Kolb received the Ph.D. degree from the University of Erlangen, Germany, in 1995. He is currently the Head of the Computer Graphics and Multimedia Systems Group, University of Siegen, Germany. His research interests include computer graphics and computer vision, including particle-based simulation and visualization, light-fields, real-time simulation, processing, and visualization of sensor data.

Michael Moeller received the Ph.D. degree from the University of Muenster, Germany, in 2012. He is currently the Head of the Computer Vision Group at the University of Siegen, Germany. His main research interests are combinations of model-based and learning-based techniques in imaging and vision along with efficient optimization algorithms to solve the underlying high dimensional minimization problems.