

# Light Field Implicit Representation for Flexible Resolution Reconstruction

Paramanand Chandramouli, Hendrik Sommerhoff, Andreas Kolb  
 E-mail: {paramanand.chandramouli, hendrik.sommerhoff, andreas.kolb}@uni-siegen.de

**Abstract**— Inspired by the recent advances in implicitly representing signals with trained neural networks, we aim to learn a continuous representation for narrow-baseline 4D light fields. We propose a novel implicit model for 4D light fields conditioned on convolutional features of a sparse set of input views. This conditioning enables our model to generalize across scenes without retraining the implicit model for each scene. Our model can be queried at continuous 4D light field coordinates, allowing joint spatial-angular super-resolution at flexible super-resolution factors. We demonstrate the flexibility of the proposed method with experiments on the tasks of view synthesis, joint spatial-angular super-resolution on real light fields. Our model outperforms current state-of-the-art baselines on these tasks, while utilizing only a fraction of run-time of the baselines. Further, our model can also be trained to be robust to varying levels of missing pixels in the input views.

## 1 INTRODUCTION

**A**CQUISITION and representation of high quality light fields is important in many diverse applications such as virtual reality, microscopy and computational photography. Light fields (LFs) are characterized using a continuous four dimensional function which represents the scene radiance along spatial and angular coordinates. As light fields dimensionally scale as  $O(n^4)$ , acquiring and storing densely-sampled LFs is expensive and challenging. Hence, several approaches have been developed to computationally reconstruct dense LFs from sparse samples [42], [50]. Recently, several deep learning-based methods have been developed for efficient reconstruction of photo-realistic novel views from sparse views [16], [44]. These approaches are often trained for a specific configuration of input observations and output resolution and offer limited flexibility.

An emerging class of neural signal representation methods referred to as implicit representations have attracted high research interest as of late [33]. These techniques provide a continuous function model for signals by parameterizing them through multi-layer perceptrons (MLPs) which are composed of deep fully connected neural networks. Apart from leading to an efficient representation, implicit representations offer the flexibility of rendering the signal value at any desired input location. Recent works, including [33], [37] have demonstrated remarkable abilities of trained neural networks in implicitly representing different classes of signals such as images, videos and 3D shapes. However, these works typically require training a separate network to represent each scene.

A few techniques have been developed in recent literature to generalize the implicit neural representations to a class of signals, without retraining from scratch for each signal. Examples of such methods include the use of meta-learning [32], [36] or training a hyper-network [33], [34] for initial network weight generation. These methods however require

further test-time optimization of weights for each signal. An alternate approach to avoid retraining is to train the implicit network by providing features from the representation space in addition to input coordinates in a supervised fashion. Such an approach has been explored in context of 3D reconstruction and shape representation in [20], [24], [25], [47] for image super-resolution [4] and scene representation [49].

In this work, we propose to learn a conditional implicit representation for 4D light fields which simultaneously can render scene radiance at continuous 4D query coordinates, while generalizing across different scenes, without the need of retraining. We develop a conditional implicit light field network (CILN) which is conditioned on very sparse set of input views to generate such implicit light field representation. Our formulation is similar to the approaches of [4], [25], [47], applied to LF data. Our model consists of a convolutional feature extractor and an implicit decoder. The CNN feature extractor embeds the input spatio-angular contextual information into the representation space. For flexibly decoding at any desired spatial resolution, the extracted features are resized to the desired resolution leading to a per-pixel latent feature representation. The 4D scene radiance is provided by an MLP decoder that is dependent on both the per-pixel features and the query spatio-angular coordinates. The use of a CNN feature extractor together with a conditioned decoder allows for scalable and robust implicit representations facilitating the reconstruction of fine-grained detail. Our approach allows super-resolution of input views simultaneously in both spatial and angular domains by flexible super-resolution factors.

In Fig. 1, we show a sample result of LF view reconstruction using our approach. The inputs to our model are the four corner views that are also of low spatial resolution. Our CILN model is able to reconstruct good quality LF views at any desired spatial and angular resolution. In Figs. 1 (c), (d) and (f), we show the output of our model at different spatial resolutions. To illustrate angular super-resolution, we show EPIs [42] in the bottom row of Fig. 1. These EPIs

• The authors are with the Department of Computer Science, University of Siegen, Siegen 57076.

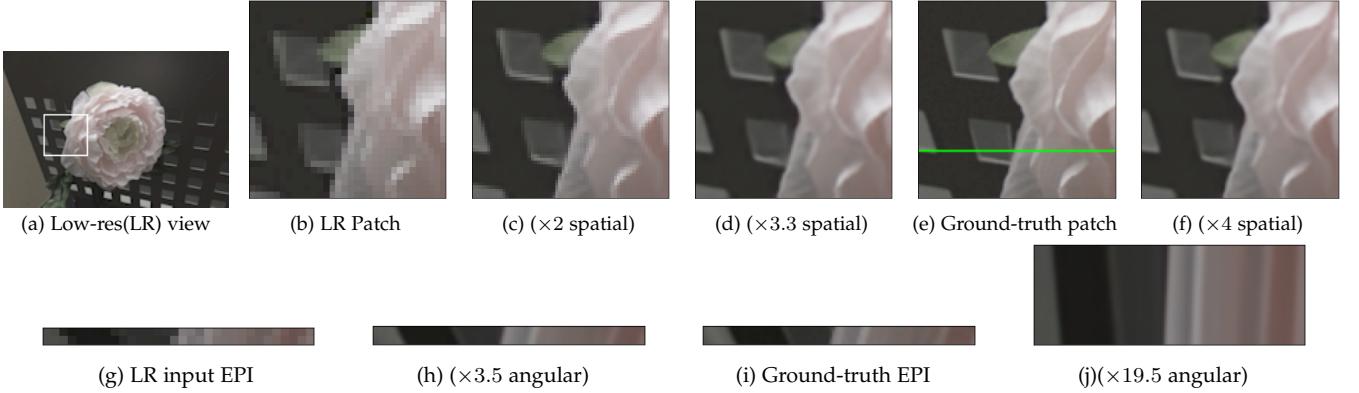


Fig. 1: Flexible resolution reconstructions from our single Conditional Implicit Light Field Network (CILN). The input to our model is a set of 4 corner views selected from the angular grid of size  $7 \times 7$ . The input views were downsampled even in spatial domain by a scale factor of 3.3. Our *single* model produces high quality reconstructions at different super-resolution factors in both spatial and angular domain. In the bottom row, we illustrate the epipolar plane image (EPI) obtained at the vertical location indicated by the green line in Fig. 1 (e). While Figs. 1 (c), (d) and (f) show flexible spatial domain reconstruction, the EPI images in Figs. 1 (h) and (j) show angular super-resolution.

show the LF plotted against horizontal spatial and horizontal angular coordinates for a specific fixed vertical location. Our reconstructions for different angular resolutions are shown in Figs. 1 (h) and (j).

Our main contributions are summarized below:

- We propose a novel conditional implicit representation for 4D light fields which can be queried at continuous coordinates, allowing *joint spatial-angular super-resolution* at *flexible* super-resolution factors.
- Our approach generalizes across scenes without requiring retraining.
- Our model achieves state of the art results on small baseline view synthesis and spatial angular super-resolution.
- Using a simple architecture, our model synthesizes high quality LF views, while utilizing only a fraction of run-time of competing view synthesis methods.
- Our model can be trained to handle challenging scenarios, where both the spatial and angular measurements are sparse, with varying levels of missing pixels.

## 2 RELATED WORK

**LF View Interpolation and Synthesis:** Recovery of dense LFs from sparse views is highly challenging. Many techniques have been proposed to tackle the problem of LF view interpolation also referred to as angular super-resolution. A good overview of existing approaches for view synthesis is available in [50]. We note that techniques have been proposed to address view synthesis for large baseline LFs such as [21]. However, we restrict our discussion and comparison to the works which consider small baseline LFs. Traditional methods for view interpolation exploit light field geometry information [42] or sparsity of LFs, for e.g. in Fourier domain [31], in learned dictionary [27], [28] or in sheared-EPI representation [39] for variational LF reconstruction. Starting from [16], several deep learning based solutions [2], [13], [14], [19], [30], [40], [43], [44], [45], [46] have been proposed

for synthesizing dense LFs. While some of these methods [19], [40], [43] employ CNNs to directly regress dense LFs from input views, others incorporate additional geometric information such as EPI structure [44], [46] or disparity-based warping [13], [14], [16], [30] into their network architecture. While EPI-based methods [44], [46] require input views on a regular grid, warping based methods can operate on irregular input views, but they typically need to be trained for each input configuration and output resolution separately. Recently, techniques [2], [13] have been proposed that can generate dense LF from a flexible set of input views. While [2] proposes to optimize the latent code of a LF generative model to fit the inputs and observation model, [13] use a plane sweep volume for disparity estimation from a flexible pattern of fixed number of input views. However, all the approaches [13], [14], [16], [19], [30], [40], [43], [44], [45] explicitly or implicitly assume the existence of correspondence across different input views and cannot handle spatial sparsity with exception to [2] which requires expensive optimization for view interpolation, and can only generate views of fixed angular resolution. Our approach can generate views of flexible spatial and angular resolution and can also be trained to handle spatial sparsity.

**LF Spatial Super-resolution:** Many approaches have been developed to overcome limited spatial resolution in LFs, some recent works include [15], [41]. While these works focus on achieving super-resolution in spatial domain only, we address a more challenging task of joint spatio-angular super-resolution by flexible factors. We note that joint spatio-angular super resolution has been attempted in the work of Meng et al. [19], for small and fixed spatial and angular super resolution factors.

**Implicit Representations:** Recent research [33], [37] has demonstrated that implicit parameterizations of continuous functions using trained multi layer perceptrons as a powerful and efficient alternative to conventional representations. Applications of implicit representations have been shown for modeling shapes and objects [9], [10], or scenes [1], [22], [33], [34], time-varying 3D geometries [23] and video and

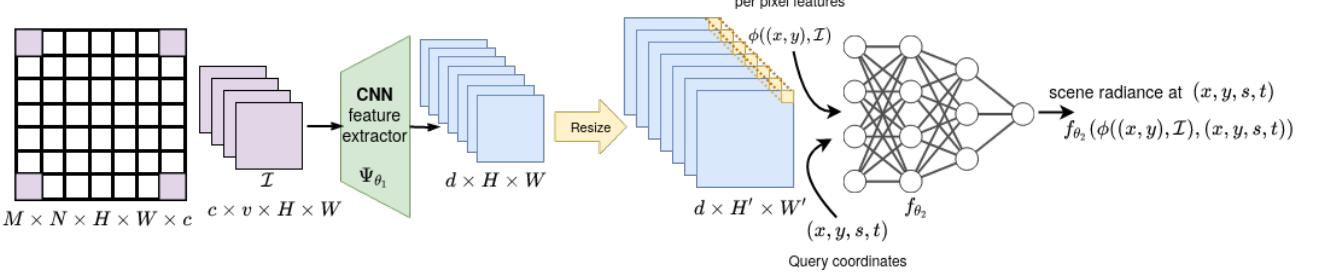


Fig. 2: Overview of Conditional Implicit Light Field Network

waveform representation [33], and more recently for 4D LFs in [6] using a novel input coordinate transformation strategy. Some of these works, including [1], [22], [23], [34] make use of differentiable rendering [38] to learn complex scene geometries using a sparse set of input views. Additionally, [1] propose to capture variations across view, time and lighting conditions by learning compact implicit neural representations with a learnable view-time geometry model with hard-coded differentiable rendering and warping. While all these methods have demonstrated excellent results in representing a signal, object or scene, they need retraining for each new instance to achieve faithful representation.

**Generalizing Implicit Representations:** To avoid retraining the implicit network from scratch for each new instance [32], [36] use meta learning to learn initial network weights for a class of shapes or signals to achieve relatively faster inference times. However, network weights need further optimization at test time for each new instance for high quality representation. An alternate approach to generalization is representing instances of signals as latent codes. Hypernetworks [11] and conditional implicit networks for 3D shapes [20], [24] use a single low dimensional latent code for each instance of signal. Closely related to such approaches are the conditional neural processes [8]. Given a set of observations and the corresponding locations, referred to as context-set, the conditional neural processes [8] and hypernetworks [11], [33], [34] aggregate latent embeddings extracted by an encoder of all inputs, which are provided to a hypernetwork to generate implicit network weights [34] or used to condition an MLP decoder by concatenating with the queried point as input [8]. Similar input concatenated latent code conditioning is exploited by the implicit networks of [20], [24]. While the approach of [20] utilizes an encoder to directly map the entire context-set to a latent embedding, the auto-decoder approach of [24] does not have an encoder and requires optimization at test time to find the optimal latent embedding corresponding to the input observations.

Though use of a single latent code for representing each instance in a class of signals, leads to a compact latent representation, such an approach leads to underfitting the context information, and therefore cannot capture fine details. Alternatively, recent approaches [5], [25], [26], [47] utilize CNNs to generate a tensor of feature embeddings which are functions of both the input coordinates and observations. The implicit network then decodes the input query point using the corresponding feature embedding, which allows representation of higher resolution details. While [4], [26],

[47], [49] operate on image data and therefore use pixel aligned implicit representations, [5], [25] can operate on the 3D point clouds, whose encodings are discretized to regular grid for further processing by CNNs.

While most conditional implicit networks have focussed on 3D shape representation and reconstruction, recent works [4], [49] have developed such representations for arbitrary super-resolution of images and scene representation respectively. Since LF views are available on a regular grid, we utilize pixel aligned representations similar to [4], [26], [47], [49]. While [4], [26], [47] learn the feature extractor, [49] uses imagenet pretrained network features. To achieve flexible spatial resolutions, we spatially upsample features similar to [4]. To capture the context across views, we learn a fused embedding at pixel level by training our feature extractor using concatenated input views, while [26], [49] average the features from each view point to obtain an aggregated embedding.

**Recovery from Missing Pixels:** Image recovery from sparse pixels has been studied since several years e.g [17], [18], [48]. Physically random pixel sampling is supported in CMOS based image sensors [29] found in hand held cameras. While traditional methods to restore missing pixels [17], [18] can operate on varying levels of sparsity, deep networks often cannot handle varying levels of corruption with a single network [7]. Deep learning based image recovery from varying levels of missing pixels has been demonstrated in [8] for a class of images, e.g. face images. However, the reconstructions are blurry due to underfitting. Similar results have not been shown for LF recovery.

### 3 PROPOSED METHOD

Let 4D light fields be denoted by the continuous function  $L(x, y, s, t)$  representing the scene radiance at spatial coordinates  $(x, y)$  and angular coordinates  $(s, t)$ . Our aim is to approximate this function implicitly using a trained neural network, which generalizes across scenes. To achieve this generalization, we condition the implicit neural network on features of a small set of input views extracted from a CNN. Fig. 2 provides an overview of the proposed *conditional implicit light field network (CILN)* framework, which consist of two parts: *feature extractor*  $\Psi_{\theta_1}$  and *scene radiance decoder*  $f_{\theta_2}$ . In practice, LF acquisition is done on a discrete grid, with fixed spatial and angular resolutions. We assume that the input  $\mathcal{I}$  to the CNN feature extractor  $\Psi_{\theta_1}$  are  $v$  views having  $c$  color channels sampled from an angular grid of size  $M \times N$  and spatial extent  $H \times W$ . To obtain a feature

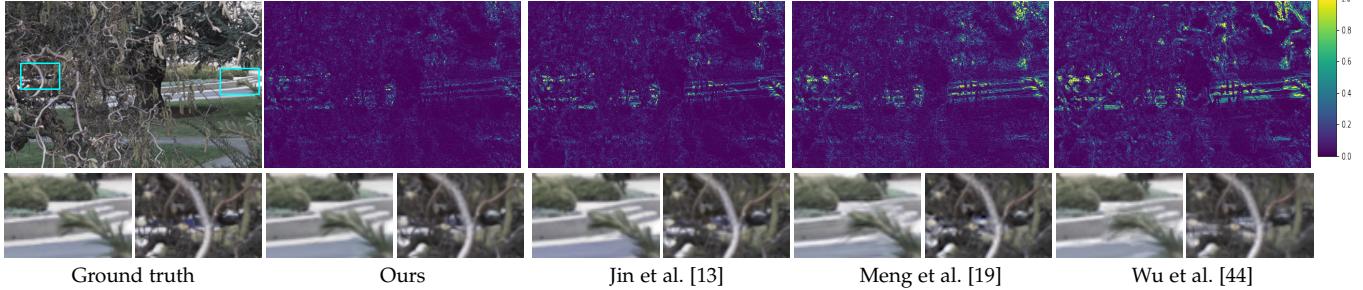


Fig. 3: Comparison of synthesized central view of the LF ‘Occlusions16’, for  $2 \times 2 \rightarrow 7 \times 7$  view upsampling. First row shows the comparison of CILN reconstruction with the ground truth and recovered views using [13], [19], [44]. Second row shows zoomed-in patches corresponding to the marked regions. Error maps are depicted with error magnified by a factor of 5.

representation, which simultaneously captures global context across views, we concatenate the input views along the color channel dimension, and train our feature extractor to learn a fused representation for the input views. We find that a simple CNN residual architecture with only two dimensional convolutions can efficiently provide feature representations. Further, the use of convolutions allows us to handle light fields of varying spatial extent, and automatically endows our output features with translational equivariance. Using appropriate padding in the convolutional layers, we obtain feature representation  $\Psi_{\theta_1}(\mathcal{I})$  of dimension  $H \times W \times d$ , where  $d$  is the per-pixel feature dimension. To facilitate reconstruction at flexible spatial resolutions, we resize the features to the desired resolution  $H' \times W'$ , the resized features are denoted by  $\phi(\mathcal{I})$ .

We utilize a multilayer perceptron (MLP)  $f_{\theta_2}$  to implicitly parameterize the scene radiance function. Given a query point with spatio-angular coordinates  $(x, y, s, t)$ , the implicit decoder  $f_{\theta}$  predicts the scene radiance (RGB values) conditioned on the per-pixel feature representation  $\phi(\mathcal{I}, (x, y))$ , i.e. the light field  $L(x, y, s, t)$  can be represented as

$$L(x, y, s, t) = f_{\theta_2}(\phi(\mathcal{I}, (x, y)), (x, y, s, t)) \quad (1)$$

**Implementation:** From a ground-truth LF patch of size  $M \times N \times H' \times W'$ , a fixed set of  $v$  views are selected as input  $\mathcal{I}$ . To facilitate flexible spatial super-resolution, we spatially downsample the input by different down sampling factors during training similar to [4]. Features are resized using bilinear interpolation. We jointly train the feature extractor  $\Psi_{\theta_1}$  and the implicit network  $f_{\theta_2}$  to reconstruct light fields from the latent embedding, by minimizing loss between the ground truth pixel values and the reconstructions. The loss function is a combination of L1 loss and EPI gradient loss [14] which encourages preservation of LF parallax structure. The feature extractor consists of four residual 2D CNN blocks [12] followed by a 1D convolutional layer. The implicit decoder is an MLP consisting of 2 hidden layers of dimension 320. We will make our code and trained models publicly available subsequently.

## 4 EXPERIMENTS

We evaluate our approach on different LF recovery tasks:

*i) View interpolation:* We train and evaluate our CILN model for view upsampling from  $2 \times 2$  views to recover  $7 \times 7$  LF.

Further, to demonstrate the flexibility of our approach, we apply the model trained for  $2 \times 2 \rightarrow 7 \times 7$  LF recovery task for reconstructing  $8 \times 8$  LFs without any retraining.

*ii) Spatial angular super-resolution:* We train CILN for flexible spatial resolutions for  $7 \times 7$  LF recovery, this model is indicated as ‘Ours<sup>†</sup>’ in the experiments. During training, we downsample input LF patches by randomly chosen scale factors ranging between 0.25 and 1. We evaluate this model for flexible spatial and angular resolutions.

*iii) LF recovery from sparse spatio-angular measurements:* We train and evaluate the CILN with varying levels of pixels missing from the  $2 \times 2$  input views for  $7 \times 7$  LF recovery. The extent of missing pixels in the input views is randomly chosen to lie in range 0% – 90% during training. This model is indicated as ‘Ours<sup>††</sup>’ in the experiments.

We used the training set of Kalantari *et al.* [16], consisting of real light fields captured using a Lytro camera. We evaluate our approach on the 30 scenes of Kalantari *et al.*’s test set, and the selected scenes (following [13]) from the ‘reflective’ and ‘occlusion’ categories of the Stanford Lytro light field archive [35].

### 4.1 View Interpolation

We quantitatively validate the performance of our approach for LF recovery from sparse input views using average PSNR and SSIM values of novel synthesized views and provide visual comparisons of the view reconstructions using error maps with respect to ground truth.

**Fixed view interpolation  $2 \times 2 \rightarrow 7 \times 7$ :** Tab. 1 provides a quantitative comparison between our method and the

Method	30scenes	Occlusions	Reflective
Wu <i>et al.</i> [44]	39.17/0.975	34.41/0.955	36.38/0.944
Meng <i>et al.</i> [19]	40.18/0.975	36.69/0.969	37.59/ 0.952
Yeung <i>et al.</i> [43]	42.77/ <b>0.986</b>	38.88/0.980	38.33/ <b>0.960</b>
Kalantari <i>et al.</i> [16]	41.40/0.982	37.25/0.972	38.09/0.953
Jin <i>et al.</i> [13]	42.75/ <b>0.986</b>	38.51/0.979	38.35/0.957
Ours	<b>42.80/0.986</b>	<b>39.36/0.981</b>	<b>39.13/0.960</b>
Ours <sup>†</sup>	41.50/0.983	38.44/0.978	38.61/0.958
Ours <sup>††</sup>	42.34/0.985	38.83/0.979	38.89/0.960

TABLE 1: Quantitative comparisons (PSNR/SSIM) of proposed CILN with the state-of-the-art view synthesis approaches for  $2 \times 2 \rightarrow 7 \times 7$  view interpolation.<sup>†</sup> indicates model trained for flexible spatial angular super-resolution. <sup>††</sup> indicates model trained for variable pixel sparsity.

following baselines : i) fully convolutional approaches of Meng et al. [19] and Yeung et al. [43] ii) warping based LF synthesis networks of Jin et al. [13] and Kalantari et al. [16] iii) deep network of Wu et al. [44] incorporating sheared EPI structures, for the three datasets considered. All the baselines are trained and tested for the task of  $2 \times 2 \rightarrow 7 \times 7$  view interpolation, with four corner views as input. We train and test our CILN for  $2 \times 2 \rightarrow 7 \times 7$  view interpolation. This model is indicated as ‘Ours’ in Tab. 1. In addition, we also compare with our CILN models trained for flexible spatial-angular super resolution (‘Ours<sup>†</sup>’) and variable pixel sparsity (‘Ours<sup>††</sup>’). Since it is very challenging to recover EPI structures using only 4 corner views, the EPI-based approach of [44] performs relatively poor, particularly in complex scenes containing occlusions and reflections. Fully convolutional approaches [19], [43] and warping based approaches [13], [16], perform better indicating the advantage of higher dimensional convolutions and geometry based warping in effectively modeling the LF structure. Our CILN trained for  $2 \times 2 \rightarrow 7 \times 7$  view interpolation outperforms all the baselines, showing marked improvement in complex scenes containing occlusions and reflections, demonstrating the advantage of the proposed approach. While our CILN can be trained to recover LFs at flexible spatial resolutions (‘Ours<sup>†</sup>’), or from variable pixel sparsity (‘Ours<sup>††</sup>’), this results in a slight degradation in performance on the fixed view interpolation task.

Fig. 3 provides the visual comparison of the synthesized central view of the scene ‘Occlusion16’. Our approach can handle complex occlusions quite well as seen in the zoomed in patches. Error maps indicate superior reconstruction of our approach, which can preserve fine grained details.

Method	30scenes	Occlusions	Reflective
Meng et al. [19]	39.25/0.970	35.72/0.964	35.62/0.945
Yeung et al. [43]	41.20/ <b>0.982</b>	<b>36.92</b> /0.971	35.74/0.946
Ours*	<b>41.30/0.982</b>	36.87/ <b>0.972</b>	<b>36.02/0.949</b>

TABLE 2: Quantitative comparisons (PSNR/SSIM) of our proposed CILN with the state-of-the-art view synthesis approaches for the task  $2 \times 2 \rightarrow 8 \times 8$  view interpolation. \* indicates our model was trained for  $7 \times 7$  view interpolation.

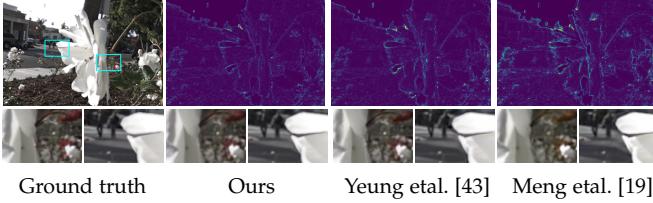


Fig. 4: Comparison of synthesized views at view location (4, 4) of the LF ‘Flower2’, for  $2 \times 2 \rightarrow 8 \times 8$  view interpolation. Top row shows the comparison of error maps of CILN reconstruction and recovered views using [19], [43] with the ground truth view. Zoomed in patches corresponding to the marked regions are shown in the second row. Error maps are depicted with error magnified by a factor of 5.

**Flexible view interpolation:** To demonstrate the ability of CILN to recover LFs of flexible angular resolution, we also evaluate  $2 \times 2 \rightarrow 8 \times 8$  LF recovery using the same model

trained for  $2 \times 2 \rightarrow 7 \times 7$  view interpolation. In Tab. 2 and Fig. 4 we provide comparison with the fully convolutional approaches of Meng et al. [19] and Yeung et al. [43] using their publicly available code and trained checkpoints. Note that the baselines have an advantage as they are specifically trained for  $8 \times 8$  view interpolation. Despite this, Tab. 2 indicates that our CILN outperforms the baselines in at least two test datasets, illustrating the benefit of the proposed approach. Visual comparison of error maps and zoomed in patches also indicates better reconstructions at the occlusion boundaries using our approach.

Method	$\times 2$	$\times 3$	$\times 4$
Yeung et al. [43]+Bicubic	35.98/0.947	32.91/0.895	31.08/0.849
Yeung et al. [43]+LIIF [4]	38.02/0.963	35.00/0.928	33.05/0.892
Ours <sup>†</sup>	<b>38.69/0.968</b>	<b>35.90/0.940</b>	<b>33.99/0.908</b>

TABLE 3: Quantitative comparisons (PSNR/SSIM) of our proposed CILN with existing approaches for  $2 \times 2 \rightarrow 8 \times 8$  angular and varying spatial upsampling on 30 scenes.

## 4.2 Flexible spatial-angular super-resolution

Feature resizing incorporated in our network architecture allows for reconstructions with flexible spatial resolutions. In our experiments, we train our CILN model (‘Ours<sup>†</sup>’) to recover  $7 \times 7$  views from variably downsampled  $2 \times 2$  input views. We have seen in Tab. 1 that this network performs view interpolation task, comparable to the baseline methods, while being slightly worse than our network trained without downsampling the inputs. To evaluate flexible spatio-angular super resolution, we test this CILN for the task of  $2 \times 2 \rightarrow 8 \times 8$  view interpolation, with flexible spatial super-resolution factors without retraining. Since there are no other existing network-based baselines for flexible spatio-angular LF upsampling, we compare our scheme by sequentially applying angular and spatial super-resolution. That is, from the  $2 \times 2$  input views, we first arrive at  $8 \times 8$  LF views using the trained model of Yeung et al. [43]. Subsequently, each of the  $8 \times 8$  views are spatially upsampled using bicubic interpolation and the state-of-the-art flexible image super-resolution scheme of LIIF [4]. Note that performing view interpolation before spatial upsampling preserves LF structure better than vice-versa.

As seen in Tab. 3, our approach achieves the best performance, indicating the benefit of joint upsampling using implicit representation. Qualitative comparisons in Fig. 5 indicate lower error and better reconstruction of fine grain detail. In contrast, results with separate upsampling in angular and spatial domains using Yeung et al. [43] and LIIF [4] have significant artifacts at the occlusion boundaries. When both spatial and angular resolution of inputs are low, it becomes highly challenging to recover fine detail. In this case, even our approach also struggles to recover fine detail as observed in results of  $\times 4$  super-resolution in Fig. 5.

## 4.3 Recovery from sparse spatio-angular inputs

Another benefit of the proposed implicit network is its ability to recover LFs from measurements which are also spatially sparse. To evaluate this, we train our CILN model (‘Ours<sup>††</sup>’) on the task of  $2 \times 2 \rightarrow 7 \times 7$  view interpolation, with

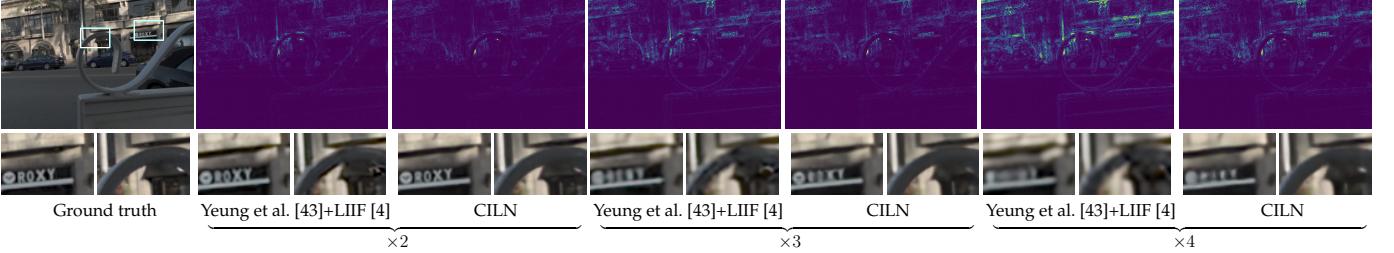


Fig. 5: Visual comparison of views at location (4, 4) for  $\times 4$  angular and  $\times 2$ ,  $\times 3$ ,  $\times 4$  spatial super-resolutions from correspondingly down sampled inputs for the scene '1586'. Shown are the ground truth view with marked regions indicating zoomed in patches, recovered views with our proposed CILN, and flexible spatial super-resolution using LIIF [4] following view interpolation using Yeung et al. [43]. Error maps are depicted with error magnified by a factor of 5.



Fig. 6: Comparison of synthesized views at angular coordinates (3,3) for the LF 'Rock', with the ground truth view. First column shows an input view to our method with missing pixels. 'p' indicates percentage of missing pixels. The columns 3 and 4 depict the reconstructed views using our CILN when 50% and 90% pixels are missing from input views. Zoomed in patches at the locations marked in the ground truth are shown.

p	30scenes	Occlusions	Reflective
0.00	42.34/0.985	38.83/0.979	38.89/0.960
0.25	41.90/0.983	38.41/0.978	38.51/0.959
0.50	41.02/0.980	37.54/0.973	37.86/0.954
0.75	38.76/0.970	35.33/0.958	36.33/0.942
0.90	34.59/0.934	31.41/0.907	33.23/0.908

TABLE 4: Average PSNR of novel views in dB for  $7 \times 7$  view synthesis using CILN trained on varying number of missing pixels 0–90%. p indicates the fraction of missing pixels.

0 – 90% of pixels, randomly dropped from input views. At test time, when there is no pixel drop, this results in only a marginal drop of performance compared to the CILN trained using clean input views (Tab. 1). The results of  $2 \times 2 \rightarrow 7 \times 7$  view interpolation for varying extents of missing pixels in the input are reported in Tab. 4. Fig. 6 depicts sample reconstruction of central view when input views have missing pixels. Even when 50% pixels are missing from inputs, CILN demonstrates a fairly faithful recovery, capturing fine details and partial occlusions. The performance degrades as expected when higher (90%) pixels are missing. The zoomed in patch in Fig. 6 shows the failure of CILN in recovering partial occlusion when 90% pixels are missing from input views. Note, that image recovery from variable levels of missing pixels is a highly challenging task, and deep networks are generally trained separately for specific amounts of degradation [7].

**Ablation study:** In Tab. 5 we investigate the effectiveness

Coordinate inputs ( $s, t$ )	$(x, y)$	MLP/ CNN	Loss	Flexible output	PSNR/SSIM
✓	✓	MLP	$L_1$ +epi-loss	✓	42.80/0.986
✓	✗	MLP	$L_1$ +epi-loss	✓	41.27/0.983
✗	✗	CNN	$L_1$ +epi-loss	✗	42.37/0.984
✓	✓	MLP	$L_1$ loss	✓	42.52/0.985

TABLE 5: Quantitative comparisons of CILN trained for  $7 \times 7$  view interpolation on 30 scenes test set with and without spatial coordinate inputs, with and without epi loss, and using implicit MLP decoder or CNN for view synthesis.

Config.	Method	30scenes	Occlusions	Reflective
	Kalantari et al. [16]	40.86/0.981	36.63/0.970	38.77/0.954
	Jin et al. [13]	42.57/0.986	39.12/0.980	40.00/0.961
	Ours	43.70/0.987	41.01/0.984	41.52/0.968
	Kalantari et al. [16]	38.54/0.973	34.83/0.958	36.82/0.950
	Jin et al. [13]	40.98/0.982	37.08/0.971	38.45/0.956
	Ours	41.74/0.983	38.57/0.977	39.60/0.960

TABLE 6: Quantitative comparisons (PSNR/SSIM) of our approach with the view synthesis approaches [13], [16] for  $7 \times 7$  view synthesis from input view configurations depicted in the first column.

of various components of our architecture and training for the task of  $2 \times 2 \rightarrow 7 \times 7$  view interpolation. In our CILN formulation, we provided both spatial coordinates ( $x, y$ ) and angular coordinates ( $s, t$ ) to the implicit network. We also evaluate the performance by CILN trained when only angular coordinates are input to CILN. We see that this results in a significant drop ( $> 1.5$ dB) in PSNR, indicating the importance of providing the 4D coordinates. Further, we also replace the MLP implicit decoder with a two layer CNN having kernel size 1 and  $49 \times 3$  output channels corresponding to the three colors channels of the  $7 \times 7$  views. Note that this does not have any coordinates as inputs and can only generate views on a fixed grid. Using such an architecture results in only a small drop in performance, showing the ability of simple 2D convolutional models in achieving competitive performance in fixed small baseline view interpolation. Our CILN model is trained using a combination of  $L_1$  loss and EPI gradient loss [14]. We also evaluate CILN trained using only  $L_1$  loss between reconstruction and ground truth. As we can see in Tab. 5, this results in a minor drop in performance compared to the original CILN trained using the combined loss.

**Discussion:** While we have mainly considered LF recovery

Algorithm	Running Time	GPU Memory
Meng et al. [19]	620 ms $\pm$ 5.37 ms	5776 MiB
Jin et al. [13]	7.52 s $\pm$ 14.4 ms	8788 MiB
Ours	237.5 ms $\pm$ 9.32 ms	10602 MiB

TABLE 7: Mean running time  $\pm$  std. dev. for  $2 \times 2$  to  $7 \times 7$  view interpolation over 10 runs. Timing and memory consumption of [13], [19] are reported for only single channel (Y channel) view synthesis, whereas the numbers reported are for view synthesis in all three RGB channels for our approach.

from  $2 \times 2$  input views, our approach can also handle irregular input views. Tab. 6 shows the results of CILN for  $7 \times 7$  LF recovery from irregular input views trained using the optimal sampling patterns from [13] for tasks of  $3 \rightarrow 7 \times 7$  and  $2 \rightarrow 7 \times 7$  LF recovery. We compare with the warping based approaches [13], [16] since they can handle irregular input view configurations. As we can see in Tab. 6, our approach can also recover high quality LFs from as low as only 2 input views which are irregularly sampled. Note that models of [13], [16] are trained for LF recovery from flexible configuration of fixed number of views. Our training with fixed pattern sampling likely gives our CILN an advantage. We will further investigate how to empower the CILN approach to handle flexible sampling patterns in future work.

**Running time:** We compare the running time and memory consumptions of our method with the methods of Jin et al. [13] and Meng et al. [19] for the task of  $2 \times 2$  to  $7 \times 7$  view recovery. This comparison was done on a machine with an Intel i9-7900X CPU @ 3.30 GHz, 128 GB RAM and an NVIDIA RTX 2080Ti GPU. For the run time experiments only, we used input LFs with a patch size of  $2 \times 2 \times 200 \times 200$ . The results in Tab. 7 show that the simple 2D convolutional architecture of our network allows a much lower computation time compared to the other two methods. Note that, our model outputs all the three color channels in the RGB space while the outputs of the [13] and [19] correspond to only the luminance component in the YCbCr space.

## 5 CONCLUSIONS

In this paper, we presented conditional implicit light field networks, a novel deep implicit representation for LFs, which generalizes across scenes. Given sparse input views, our CILN predicts the scene radiance at any queried point in the spatial and angular dimensions. Our framework achieves this by propagating the input context through a convolutional neural network, to provide pixel level local fused embeddings. Our implicit network exploits these local embeddings to capture fine-grained details and generate a photorealistic LF reconstruction. Qualitative and quantitative experiments validate that our CILN can provide reconstructions outperforming recent state of the art approaches for LF view synthesis on real scenes. Our CILN can generate LF views at arbitrary spatio-angular resolutions clearly demonstrating our flexibility. Further, CILN can also be trained to be robust to varying levels of spatial sparsity with a single model. Future work may include extending such implicit representations to much larger baseline light fields.

## REFERENCES

- [1] M. Bemana, K. Myszkowski, H. Seidel, and T. Ritschel. X-fields: Implicit neural view-, light- and time-image interpolation. *ACM Trans. Graphics*, 39(6), 2020.
- [2] P. Chandramouli, K. V. Gandikota, A. Gorlitz, A. Kolb, and M. Moeller. A generative model for generic light field reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.
- [3] S. Chaudhury and H. Roy. Can fully convolutional networks perform well for general image restoration problems? In *IAPR International Conference on Machine Vision Applications*, IEEE, 2017.
- [4] Y. Chen, S. Liu, and X. Wang. Learning continuous image representation with local implicit image function. In *Proc. CVPR*, 2021.
- [5] J. Chibane, T. Alldieck, and G. Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proc. CVPR*, June 2020.
- [6] B.Y. Feng and A. Varshney. SIGNET: Efficient neural representation for light fields. In *Proc. ICCV*, 2021.
- [7] R. Gao and K. Grauman. On-demand learning for deep image restoration. In *Proc. ICCV*, pages 1086–1095, 2017.
- [8] M. Garnelo, D. Rosenbaum, C. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. Rezende, and SM A. Eslami. Conditional neural processes. In *Proc. ICML*, pages 1704–1713. PMLR, 2018.
- [9] K. Genova, F. Cole, D. Vlasic, A. Sarna, W. T. Freeman, and T. Funkhouser. Learning shape templates with structured implicit functions. In *Proc. ICCV*, pages 7154–7164, 2019.
- [10] A. Gropp, L. Yariv, N. Haim, M. Atzman, and Y. Lipman. Implicit geometric regularization for learning shapes. In *Proc. ICML*, 2020.
- [11] D. Ha, A. Dai, and Q. V Le. Hypernetworks. *Proc. ICLR*, 2017.
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.
- [13] J. Jin, J. Hou, J. Chen, H. Zeng, S. Kwong, and J. Yu. Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.
- [14] J. Jin, J. Hou, H. Yuan, and S. Kwong. Learning light field angular super-resolution via a geometry-aware network. In *Proc. AAAI*, 2020.
- [15] J. Jin, J. Hou, J. Chen, and S. Kwong. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In *Proc. CVPR*, 2020.
- [16] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM Trans. Graphics*, 2016.
- [17] M. Li, J. Liu, Z. Xiong, X. Sun, and Z. Guo. Marlow: A joint multi-planar autoregressive and low-rank approach for image completion. In *Proc. ECCV*, Springer, 2016.
- [18] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):208–220, 2012.
- [19] N. Meng, H. K.-H. So, X. Sun, and E. Lam. High-dimensional dense residual convolutional neural network for light field reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.
- [20] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. CVPR*, pages 4460–4470, 2019.
- [21] 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 Trans. Graphics*, 38(4):1–14, 2019.
- [22] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. ECCV*, 2020.
- [23] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *Proc. ICCV*, 2019.
- [24] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. CVPR*, pages 165–174, 2019.
- [25] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger. Convolutional occupancy networks. In *Proc. ECCV*, 2020.
- [26] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proc. ICCV*, 2019.
- [27] D. C. Schedl, C. Birkbauer, and O. Bimber. Directional super-resolution by means of coded sampling and guided upsampling. In *Proc. ICCP*, pages 1–10. IEEE, 2015.
- [28] D. C. Schedl, C. Birkbauer, and O. Bimber. Optimized sampling for view interpolation in light fields using local dictionaries. *Computer Vision and Image Understanding*, 168:93–103, 2018.

- [29] D. Scheffer, B. Dierickx, and G. Meynants. Random addressable 2048/spl times/2048 active pixel image sensor. *IEEE Trans. Electron Devices*, 44(10):1716–1720, 1997.
- [30] J. Shi, X. Jiang, and C. Guillemot. Learning fused pixel and feature-based view reconstructions for light fields. In *Proc. CVPR*, 2020.
- [31] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand. Light field reconstruction using sparsity in the continuous fourier domain. *ACM Trans. Graphics*, 34(1):1–13, 2014.
- [32] V. Sitzmann, E. Chan, R. Tucker, N. Snavely, and G. Wetzstein. Metasdf: Meta-learning signed distance functions. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Proc. Neurips*, pages 10136–10147, 2020.
- [33] V. Sitzmann, J. N.P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. Neurips*, 2020.
- [34] V. Sitzmann, M. Zollhoefer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Proc. Neurips*, 2019.
- [35] A. SunderRaj, M. Lowney, R. Shah, and G. Wetzstein. Stanford lytro light field archive. <http://lightfields.stanford.edu/LF2016.html>, 2016.
- [36] M. Tancik, B. Mildenhall, T. Wang, D. Schmidt, P. P. Srinivasan, J. T. Barron, and R. Ng. Learned initializations for optimizing coordinate-based neural representations. *arXiv preprint arXiv:2012.02189*, 2020.
- [37] M. Tancik, Pratul P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singh, R. Ramamoorthi, J. T. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Proc. Neurips*, 2020.
- [38] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. M. Brualla, T. Simon, J. Saragih, and M. Nießner. State of the art on neural rendering. *Proc. Eurographics*, 2020.
- [39] S. Vagharshakyan, R. Bregovic, and A. Gotchev. Light field reconstruction using shearlet transform. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(1):133–147, 2017.
- [40] 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 *Proc. ECCV*, 2018.
- [41] Y. Wang, J. Yang, Longguang Wang, X. Ying, T. Wu, W. An, and Yulan Guo. Light field image super-resolution using deformable convolution. *IEEE Trans. Image Processing*, 30:1057–1071, 2021.
- [42] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):606–619, 2013.
- [43] H. W. F. Yeung, J. Hou, J. Chen, Y. Y. Chung, and X. Chen. Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In *Proc. ECCV*, 2018.
- [44] G. Wu, Y. Liu, Q. Dai, and T. Chai. Learning sheared epi structure for light field reconstruction. *IEEE Trans. Image Processing*, 2019.
- [45] G. Wu, Y. Liu, L. Fang, and T. Chai. Spatial-angular attention network for light field reconstruction. *arXiv preprint arXiv:2007.02252*, 2020.
- [46] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai. Light field reconstruction using convolutional network on epi and extended applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7):1681–1694, 2019.
- [47] Q. Xu, W. Wang, D. Ceylan, R. Mech, and U. Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Proc. Neurips*, 2019.
- [48] R. Yeh, C. Chen, T. Y. Lim, M. H. Johnson, and M. N. Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2016.
- [49] A. Yu, V. Ye, M. Tancik and A. Kanazawa pixelnerf: Neural radiance fields from one or few images. In *Proc. CVPR*, 2021.
- [50] D. Yue, M. S. Khan Gul, M. Bätz, J. Keinert, and R. Mantiuk. A benchmark of light field view interpolation methods. In *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2020.