

Learning Occlusion-Aware View Synthesis for Light Fields

Julia Navarro and Neus Sabater

Abstract—In this work, we present a novel learning-based approach to synthesize new views of a light field image. In particular, given the four corner views of a light field, the presented method estimates any in-between view. We use three sequential convolutional neural networks for feature extraction, scene geometry estimation and view selection. Compared to state-of-the-art approaches, in order to handle occlusions we propose to estimate a different disparity map per view. Jointly with the view selection network, this strategy shows to be the most important to have proper reconstructions near object boundaries. Ablation studies and comparison against the state of the art on Lytro light fields show the superior performance of the proposed method. Furthermore, the method is adapted and tested on light fields with wide baselines acquired with a camera array and, in spite of having to deal with large occluded areas, the proposed approach yields very promising results.

Index Terms—Light field image, new view synthesis, convolutional neural networks

1 INTRODUCTION

LIGHT field imaging has recently gained importance due to the additional information that provides of the scene. Contrary to conventional 2D images that at each point capture the sum of all light rays coming from different angles, the 4D light field image captures the whole light information. A light field image can be considered as a collection of 2D images taken from different viewpoints that are arranged on a regular grid.

Plenoptic cameras such as Lytro [1] or camera arrays [2], [3] are among the different devices that can be used for the acquisition of these images. In the first case, given that the sensor resolution is limited, the additional information given from the different viewpoints comes at the cost of an important decrease in spatial resolution, compared to traditional cameras. Plenoptic cameras usually offer high angular resolutions (14×14 views for Lytro Illum) with small baselines. On the other hand, camera arrays do not suffer from low spatial resolution but capturing a large number of views would be costly, and generally they capture sparse light fields with wide baselines. In addition, current smartphones also capture light fields using several cameras. However, they cannot provide high angular resolutions since it is not possible to have a large number of cameras in a cellphone.

Then, it is interesting to study the problem of new view synthesis for light fields. That is, the generation of images from novel viewpoints. With new view synthesis methods plenoptic cameras could be built to capture light fields with smaller angular resolution and thus provide higher spatial resolutions. Also, camera arrays and cellphones could increase the number of views using view synthesis techniques. Besides, the generation of novel views would permit to navigate smoothly between the different images and be used

for applications such as virtual reality [4].

Over the last years, deep learning has had a great success in computer vision and image processing tasks. Indeed, deep learning has proved to be competitive with respect to traditional approaches for different problems such as stereo [5], [6], optical flow [7], denoising [8] and super-resolution [9]. Furthermore, it has recently been applied to light field images for super-resolution [10], depth estimation [11] or separation into diffuse and specular intrinsic components [12].

Inspired by recent work on new view synthesis using deep learning [13], [14], [15], we propose a novel learning-based solution to synthesize views of a light field image. Particularly, given de four corner views, we reconstruct any view in between. The approach is designed for plenoptic light fields captured with the Lytro Illum camera. Moreover, we adapt the model for wide baselines and very promising results are obtained in the case of light fields captured with a camera rig, which have larger occluded regions.

We divide the problem of view synthesis into feature extraction, disparity estimation and view selection and use three sequential convolutional neural networks. Disparity is estimated between the virtual novel view and each corner view. In contrast to the recent approach from Kalantari et al. [14], in order to handle occlusions we propose to estimate a different disparity map for each corner image. The selection network detects occluded parts and discards them to reconstruct the novel view. This results in accurate reconstructions near object boundaries and occlusions, while the method in [14] produces blurred results at these regions. Srinivasan et al. [15] reconstruct the 4D light field from the center view. While their problem is more challenging than ours, they work with images with simple and similar geometry, being unable to deal with more complex scenes. In spite of having been trained on the same dataset, our approach performs properly on different scenes. Flynn et al. [13] cope with wide baseline images by providing to the networks a plane sweep volume built from the input views.

- J. Navarro is with Universitat de les Illes Balears, Palma, 07122 Spain.
E-mail: julia.navarro@uib.es
- N. Sabater is with Technicolor R&I, Cesson-Sévigné, 35576 France.
E-mail: neus.sabater@technicolor.com

While this is memory and time consuming (it takes minutes to synthesize a novel view), our method takes few seconds to predict the novel image directly from the input views.

Although the purpose of the work is view synthesis, the presented method is able to estimate disparity. Since we only need a large collection of light fields for training and any ground truth depth is needed, it learns disparity in an unsupervised manner. Furthermore, the disparity estimation carried out by our method is competitive with respect to the state of the art.

The work is organized as follows. Section 2 reviews the state of the art on view synthesis. In Section 3, we present the proposed new view synthesis model. In Section 4 we evaluate the presented method with extensive experiments and comparisons. Finally, Section 5 concludes the paper.

2 PREVIOUS WORK

In this section, we review the state of the art on new view synthesis. We divide these methods into non-learning and learning-based approaches. In addition, as the problem is closely related to video frame interpolation, we also introduce recent methods tackling this problem.

2.1 Non-Learning Approaches

Traditional methods generate novel views of scenes and objects from an arbitrary collection of images from the scene, which is known as image-based rendering (IBR). Generally, these methods first predict the scene geometry and then generate the novel image from the warped views. Chaurasia et al. [16] proposed a depth-synthesis approach using graphs operating over an over-segmentation of the input views. Goesele et al. [17] introduced ambient point clouds to represent areas with uncertain depth. Other methods estimate the novel image without explicitly estimating geometry. Fitzgibbon et al. [18] avoided the explicit depth computation and used image-based priors. Shechtman et al. [19] used a patch-based optimization framework.

Among the non-learning view synthesis methods for light field images we find the variational model from Wanner and Goldluecke [20]. Given the disparity maps at the input views, the energy functional penalizes deviations between each warped input onto the novel position and the unknown view. This term incorporates a mask to account for occlusions. Also, a smoothness term for the novel image using total variation is included. Shi et al. [21] work in the continuous Fourier domain to reconstruct dense light fields from a 1D set of viewpoints. Zhang et al. [22] proposed a phase-based approach to reconstruct a whole light field from a stereo pair with disparities smaller than five pixels. Penner and Zhang [23] generate novel views of plenoptic light fields and camera arrays by means of a soft 3D model of the scene geometry.

2.2 Learning-based Methods

More recent methods make use of convolutional neural networks (CNN) to model the problem. Yoon et al. [10] jointly model spatial and angular light field super resolution with a CNN. The spatially upsampled result is the input to the angular super resolution network, which consists of

a single CNN. Wu et al. [24] reconstruct any view of the light field given a sparse set of views by using a CNN on epipolar plane images. Similarly, Wang et al. [25] combine 2D and 3D convolutions applied on epipolar plane images to reconstruct the entire light field. From the four corner views of a light field, Kalantari et al. [14] propose two CNN to synthesize any view in between. They manually extract features by first warping the input images at different disparity levels and then computing mean and variance at each level. Given these features, the first network computes one disparity map for the unknown view, which is used to warp each corner image. The four warpings are combined through another CNN which outputs the predicted view. Srinivasan et al. [15] aim at reconstructing the whole light field given just the center view. A first network estimates a 4D depth map from the input image. These maps are used to warp the center view and obtain an initial estimate of the 4D light field, which is further refined through a residual network. The method is trained on images of flowers, all of them sharing similar geometry, and it fails when testing on more complex scenes.

Flynn et al. [13] deal with wide-baseline images by building a plane sweep volume. This volume is the input to two different networks, one that outputs for each pixel and depth the probability of that pixel having that depth, and the other generates a color image at each depth plane. The point-wise product between probabilities and color images provide the novel view. A drawback of this approach is the need to build the plane sweep volume, which is memory and time consuming. Indeed, the authors have to generate images in small patches to save memory and it takes 12 minutes to synthesize a 512×512 image. Plane sweep volumes are also used by Zhou et al. [26], where the authors develop a method for view extrapolation given two images with small baseline, using an encoder-decoder architecture. Built upon this method, Srinivasan et al. [27] further extend the possible lateral movement and improve reconstructions at disocclusions. The recent method from Mildenhall et al. [28] also uses plane sweep volumes to synthesize novel views given an irregular grid of input images from the scene.

2.3 Video Frame Interpolation

Given two frames of a video, frame interpolation consists of predicting frames at novel time instants. Most approaches first estimate optical flow to warp the input frames to the target one and then proceed to combine these warpings.

Liu et al. [29] proposed Deep Voxel Flow, a multiscale frame interpolation method. At three different scales, they compute the optical flow and a confidence map using encoder-decoder networks. The results from these scales are combined using two convolutional layers. The output vector field is used to warp the input images and both warpings are combined using the confidence map. Amersfoort et al. [30] estimate optical flow and confidence in a coarse-to-fine scheme to deal with large displacements. At each scale the method estimates a residual to refine both the flow estimation and confidence. The finest optical flow computation is at half resolution, and it is upsampled to warp the frames at full resolution and generate the novel one. This result is further refined through a CNN. Niklaus and

TABLE 1
Networks architectures.

	Name	k	r	In	Out	Act. f.	BN
Features CNN	input				5		
	conv0	3×3		5	32	ELU	✓
	conv1	3×3		32	32	ELU	✓
	conv2	3×3		32	32	ELU	✓
	conv3	3×3		32	32	ELU	✓
	conv4	3×3		32	32	ELU	✓
	conv4 = conv2 + conv4						
	pool0	16×16		32	32	avg. (conv4)	
Disparity CNN	pool1	8×8		32	32		
	concatenate [conv2, conv4, pool0, pool1]						
	conv5	3×3		128	32	ELU	✓
	input				130		
	conv0	3×3	2	130	128	ELU	✓
	conv1	3×3	4	128	128	ELU	✓
	conv2	3×3	8	128	128	ELU	✓
	conv3	3×3	16	128	128	ELU	✓
Selection CNN	conv4	3×3		128	64	ELU	✓
	conv5	3×3		64	64	ELU	✓
	conv6	3×3		64	4	tanh	
	$d_{\max} \cdot \text{conv6}$						
	input				18		
	conv0	3×3		18	64	ELU	✓
	conv1	3×3		64	128	ELU	✓
	conv2	3×3		128	128	ELU	✓
	conv3	3×3		128	128	ELU	✓
	conv4	3×3		128	64	ELU	✓
	conv5	3×3		64	32	ELU	✓
	conv6	3×3		32	4	tanh	
	Softmax with learned β						

Liu [31] use an existing deep learning method to compute forward and backward optical flows. These are used to warp the input frames as well as the features provided by the first layer of the ResNet18 [32]. The four warpings are the input to a network with a GridNet architecture [33]. Jiang et al. [34] use an encoder-decoder network to predict forward and backward flows. These are used to warp the frames to the desired time instant and input views, warpings and optical flows are introduced into another encoder-decoder network to refine the optical flow. This network outputs the refined flow, jointly with a confidence mask. Then, the input frames are warped with these refined flows and combined according to the confidences.

3 PROPOSED METHOD FOR VIEW SYNTHESIS

Let $\Omega \subset \mathbb{R}^2$ be an open bounded domain, usually a rectangle in \mathbb{R}^2 , and let us consider a light field image with $(N + 1) \times (N + 1)$ views, with $N \in \mathbb{N}, N \geq 2$. Let us denote by $I_{p,q} : \Omega \rightarrow \mathbb{R}^3$ the view at the angular position (p, q) , with $p, q \in [0, N]$ and with the $(0, 0)$ image being the one at the top-left corner.

Given the four corner images $I_{0,0}, I_{0,N}, I_{N,0}$ and $I_{N,N}$ and the angular coordinates (p, q) of any in-between view, the goal is to estimate the view $I_{p,q}$. That is, we aim at finding a function f such that

$$\hat{I}_{p,q} = f(p, q, I_{0,0}, I_{0,N}, I_{N,0}, I_{N,N}), \quad (1)$$

with $\hat{I}_{p,q}$ being the estimated view at position (p, q) .

We model f by using convolutional neural networks. One option would be to consider f as a single network that from the four corner views and the coordinates of the novel

position directly outputs the predicted view. However, as pointed out in [14], [15], the relation between input and output is too complex to be modeled by just a single network. A proof of that is later shown in Section 4.

3.1 Proposed Model

We split the problem into feature extraction, disparity estimation and view selection and use three different convolutional neural networks, one for each purpose. Features extracted from four input images are concatenated and used to estimate disparity. Then, input views are warped according to this disparity and four selection masks that will serve to perform a weighted average of the four warpings are estimated.

The three networks additionally receive as input the coordinates (p, q) of the novel view. In order to provide these coordinates to the convolutional networks, we consider images $P, Q : \Omega \rightarrow \mathbb{R}$ such that

$$\begin{aligned} P(x, y) &= p, & \forall (x, y) \in \Omega, \\ Q(x, y) &= q, & \forall (x, y) \in \Omega. \end{aligned} \quad (2)$$

In the following we detail each stage of our algorithm.

3.1.0.1 Features CNN: Compared to [14] that extracts features manually, we use a convolutional neural network for this purpose. The features extraction network (f_e) is applied independently to each input image to compute a feature volume with 32 channels for each one of the four input views. These features should not depend on the image being processed and therefore weights are shared across all views. This network also receives as input the images P and Q , which are concatenated to the considered image along the channel dimension, resulting in an input volume with 5 channels.

The network f_e consists of a sequence of five convolutional layers with 3×3 kernels, including one residual block [32]. Average poolings with kernels 16×16 and 8×8 are then used to extract features at different scales, providing the network of more global information. Features from different layers are concatenated and finally fused with 3×3 convolutions. All convolutional layers are followed by an ELU activation and batch normalization [35]. This architecture is a simplified version of the feature extraction stage proposed in [6].

Let $F_{i,j} = f_e(P, Q, I_{i,j})$ be the computed feature volume for image $I_{i,j}$, for $i, j \in \{0, N\}$. Then, the four volumes are concatenated,

$$\mathbf{F} = (F_{0,0}, F_{0,N}, F_{N,0}, F_{N,N}), \quad (3)$$

and this 128-channel volume \mathbf{F} is the input to the next stage.

3.1.0.2 Disparity CNN: We assume that the views of the light field are arranged on a regular grid. Then, horizontal and vertical disparities are the same for consecutive views and thus the same estimated map is used in both components. For the same reason, disparities between each corner view and the virtual view are the same and one common map for the four images should be enough. In practice, however, the matching problem is not defined at occluded areas and, since occluded pixels are different depending on the view, it results in different disparity maps. Therefore, in contrast to [14], we let the network to estimate four different

disparity maps $d_{i,j}$ depicting the displacement between $I_{i,j}$ and the virtual view $\hat{I}_{p,q}$, for $i, j \in \{0, N\}$. In Section 4 we show the advantages of using this strategy.

The disparity maps $\mathbf{d} = (d_{0,0}, d_{0,N}, d_{N,0}, d_{N,N})$ are computed from the angular position of the novel view and the four feature volumes, \mathbf{F} , through network f_d ,

$$\mathbf{d} = f_d(P, Q, \mathbf{F}). \quad (4)$$

This network consists of seven convolutional layers, all of them with a filter size of 3×3 . The first four ones use dilated convolutions at rates 2, 4, 8 and 16, respectively. The use of dilated convolutions permits to combine features at different resolutions and provide the network with more context. All layers but the last one use an ELU activation function and batch normalization. Last layer uses the hyperbolic tangent as activation function and no batch normalization is applied. The tanh rescales the output into the range $[-1, 1]$. Then, the output disparity is multiplied by a constant d_{\max} , which is the maximum allowed disparity magnitude. This way the output disparity will be in the range $[-d_{\max}, d_{\max}]$. For Lytro images, this value is set to $d_{\max} = 4$.

3.1.0.3 Image Warping: The estimated disparity is used to warp each corner view in order to have them registered with the virtual one. Let $I_{i,j}^w$ denote the warped image for view $I_{i,j}$. Then, for all $i, j \in \{0, N\}$,

$$I_{i,j}^w(x, y) = I_{i,j}(x + (i - p)d_{i,j}, y + (j - q)d_{i,j}), \quad (5)$$

where $d_{i,j}$ is evaluated at pixel (x, y) . Warped images and disparity maps are concatenated to form the volume \mathbf{W} ,

$$\mathbf{W} = (I_{0,0}^w, I_{0,N}^w, I_{N,0}^w, I_{N,N}^w). \quad (6)$$

This 12-channel volume \mathbf{W} , the depth maps \mathbf{d} and images P and Q are the input to the selection network.

3.1.0.4 Selection CNN: The task of the selection network (f_s) is to determine the contribution of each warped image $I_{i,j}^w$ to the final result. This will be achieved by computing four selection masks $(m_{0,0}, m_{0,N}, m_{N,0}, m_{N,N}) = f_s(P, Q, \mathbf{W}, \mathbf{d})$ such that $m_{i,j}(x, y) \in [0, 1]$, for all $i, j \in \{0, N\}$, and

$$\sum_{i,j \in \{0,N\}} m_{i,j}(x, y) = 1, \quad \forall (x, y) \in \Omega. \quad (7)$$

Then, the predicted view is computed as a weighted average of the four warped images using as weights these selection masks,

$$\hat{I}_{p,q}(x, y) = \sum_{i,j \in \{0,N\}} m_{i,j}(x, y) I_{i,j}^w(x, y). \quad (8)$$

The selection network f_s consists of seven convolutional layers with 3×3 filters. All layers but the last one are followed by an ELU and batch normalization. At the last layer we use tanh and do not use batch normalization. Besides, at the last layer we also apply a softmax normalization along views,

$$\sigma_\beta(v_i(\mathbf{x})) = \frac{e^{\beta v_i(\mathbf{x})}}{\sum_{i=1}^4 e^{\beta v_i(\mathbf{x})}}, \quad \forall i \in \{1, 2, 3, 4\}, \quad (9)$$

with $\mathbf{x} = (x, y)$ and v_i being channel i of the conv6 layer. With the softmax we ensure that the sum of the selection weights over the four views equals one at each

pixel. Moreover, we let the network to learn the parameter β . High values of this parameter encourage the network to select a single view, which is important at those areas that are only visible in one of the four images. The network has to be able to detect which regions of the novel view are also visible in the four corner ones. With these masks we discard inaccuracies in the warped images coming from occluded pixels. After training the network, the learned value is $\beta = 8.01$.

Table 1 details the three presented networks. In the table, labels *In* and *Out* correspond to the number of channels of input and output volumes, respectively. BN denotes batch normalization [35], k is the kernel size and r the dilation rate, which equals one when nothing specified. Moreover, zero padding is applied to all layers to maintain spatial dimensions.

3.2 Loss Function for Network Optimization

The loss energy function proposed to train the model consists of two terms. The first term penalizes deviations between the reconstructed view and ground truth image:

$$E_d = \|I_{p,q} - \hat{I}_{p,q}\|_1. \quad (10)$$

To better preserve image textures, the second proposed term additionally imposes the output image to have similar spatial gradients to the ground truth:

$$E_g = \|\nabla I_{p,q} - \nabla \hat{I}_{p,q}\|_1. \quad (11)$$

Then, the proposed loss function writes as

$$E = E_d + \lambda_g E_g, \quad (12)$$

where we experimentally set $\lambda_g = 0.5$. In Section 4 we evaluate different configurations for this training loss.

Another term we could have included is one that enforces consistency between different disparity maps, similar to [15]. However, disparity maps should not be equal at occluded regions and, since we do not know these occlusions beforehand, we do not impose any constraint.

3.3 Training Details

The model has been implemented using TensorFlow [36]. We train the networks on Lytro light fields which have a spatial resolution of 540×372 and an angular one of 14×14 , from which we select a centred 7×7 array of views. The four corner views of these 7×7 light fields are the inputs to our method.

At each training iteration, we randomly select the angular coordinates at integer positions $p, q \in \mathbb{Z} \cap [0, 6]$, excluding the corner views. The output is compared at each iteration to the ground truth view by means of the loss function presented in Equation (12). We randomly extract 192×192 patches from the training images to train the model. The network is optimized using the ADAM solver [37] with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e - 08$, a learning rate of 0.001 and a batch size of 3. Weights are initialized randomly using the Xavier method [38] and the softmax β is initialized to 1. The method converges after 300k iterations and it approximately takes 1 day and 20 hours on a GeForce GTX 1080 Ti GPU. At test time, it takes less than 2 seconds to synthesize a 540×372 image.

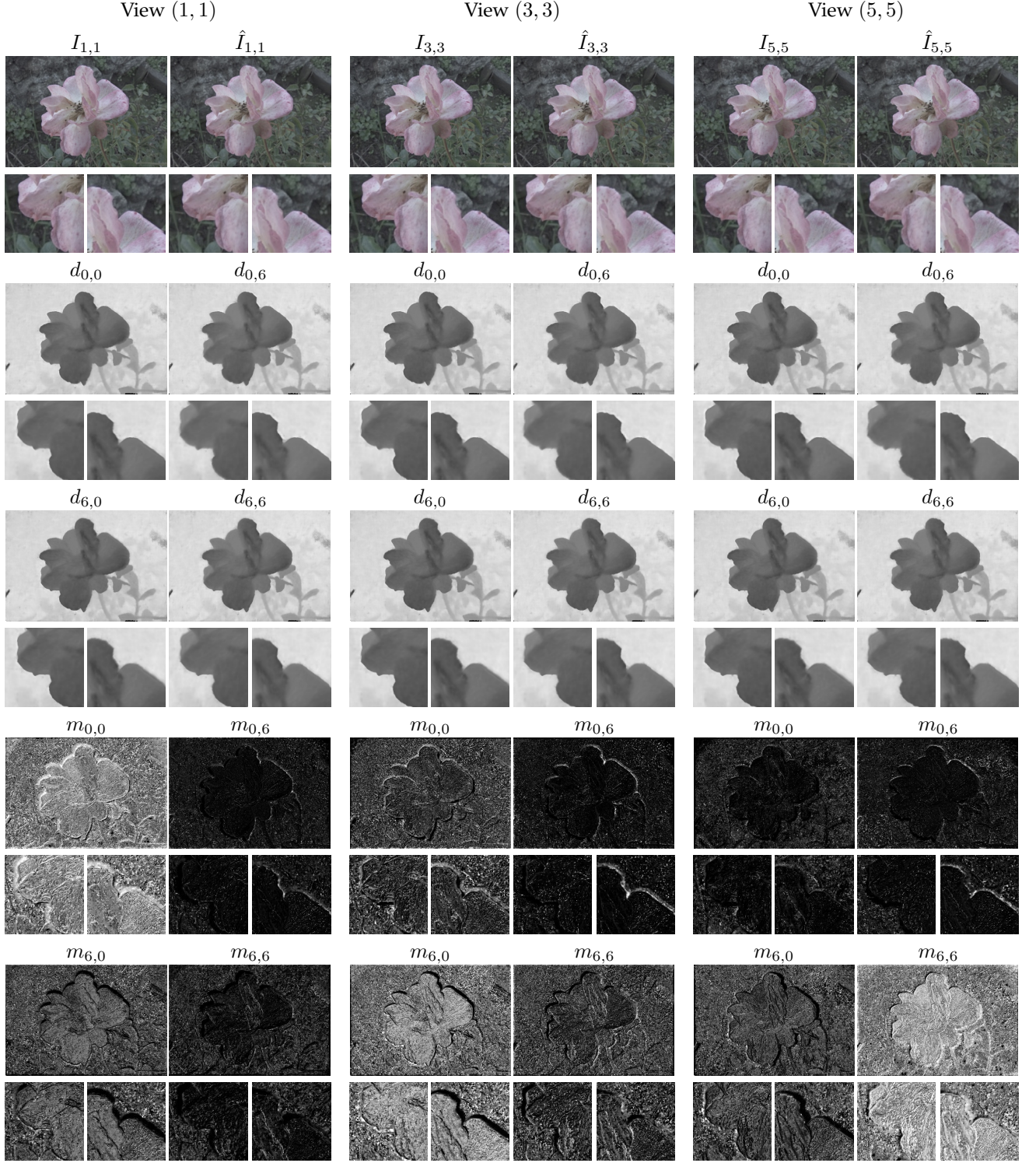


Fig. 1. Visual results for three different input angular positions for the novel view. We display the ground truth and estimated view (top), the predicted disparity maps for each corner image (middle) and the four selection masks (bottom). We encourage the reader to look at the electronic version of this paper to better see the details.

4 EXPERIMENTS

In this section, we evaluate the performance of the proposed method. First, we assess the different components included in our approach by means of several experiments. Then, we compare the obtained results against state-of-the-art methods for light field view synthesis. Finally, the method is adapted and tested on light fields acquired with an array

of cameras.

The quantitative evaluation reported during this section is in terms of the mean absolute error, which is multiplied by 100 for images in the range $[0, 1]$ (MAE), the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) [39]. Unless otherwise stated, the reported metrics are averaged over all possible in-between viewpoints

TABLE 2
Analysis of different terms in the loss function.

			<i>Flowers</i>			<i>Diverse</i>		
E_d	E_g	E_w	MAE	PSNR	SSIM	MAE	PSNR	SSIM
✓			0.887	38.07	0.9770	0.820	37.82	0.9834
✓	✓		0.879	38.28	0.9778	0.799	38.12	0.9848
✓	✓	✓	0.934	37.74	0.9757	0.820	37.87	0.9846

at integer positions, excluding the input corner ones, and over the whole test set under consideration. Besides, when nothing specified, all presented visual results correspond to the center view, which has angular coordinates (3, 3). Finally, for the sake of simplicity, in some cases we just display one disparity map as the disparity estimated by the proposed method, which actually corresponds to $d_{0,0}$.

4.1 Datasets

We used two different datasets for training and testing the proposed method. On the one hand, the dataset from Srinivasan et al. [15], which consists of 3343 images of flowers captured with a Lytro Illum camera. We randomly divided it into 3243 images for training and 100 for testing the model. On the other hand, the dataset from Kalantari et al. [14]. It contains 100 light fields for training and 30 for testing. They are mostly outdoor images from diverse scenarios captured with the Lytro Illum. When reading the images, we apply a gamma correction with $\gamma = 0.4$ to both datasets. We denote as *Flowers* the dataset from Srinivasan et al. [15] and as *Diverse* the one from Kalantari et al. [14]. For the experiments in this section, when nothing specified, the used dataset for training is *Flowers*.

4.2 Visual Results

Figure 1 visually illustrates the performance of the proposed model on one example of the *Flowers* test set and for three different angular coordinates for the novel view. As it can be seen in the figure, disparity maps are sharp at all depth discontinuities but they are more blurred at occlusions. At occluded regions, the warped views will be inaccurate. However, with the selection network we are able to discard occluded pixels. Occluded parts are equal to zero in the selection masks and more weight is given to the areas that are visible in only one view. Also, we can appreciate how the selection network has a preference on choosing the warped view whose angular position is closest to the novel one. This occurs because Lytro light fields present changes in color between views and the closer the viewpoints are, the more similar color the images have.

4.3 Analysis of the Loss Function

Table 2 reports evaluation metrics for three different configurations of the training loss. First, the model trained with the only use of the reconstruction error term E_d (10). Second, using the proposed E_d and the gradients difference term E_g (11). Third, apart from E_d and E_g , we additionally include a term E_w that enforces each disparity estimation to be consistent with the warped views. That is,

$$E_w = \frac{1}{4} \sum_{i,j \in \{0,N\}} \|I_{p,q} - I_{i,j}^w\|_1. \quad (13)$$

TABLE 3
Comparison against one single network, the use of just one disparity map and without the use of the features network.

			<i>Flowers</i>			<i>Diverse</i>		
Method	Param.	Time	MAE	PSNR	SSIM	MAE	PSNR	SSIM
1 CNN	1.66 M	1.40 s	10.69	24.54	0.9488	11.62	23.86	0.9493
1 disp.	1.27 M	1.91 s	0.931	37.76	0.9750	1.030	36.03	0.9732
w/o f_s	1.27 M	1.82 s	0.931	37.81	0.9757	1.080	35.66	0.9703
Proposed	1.27 M	1.89 s	0.879	38.28	0.9778	0.799	38.12	0.9848

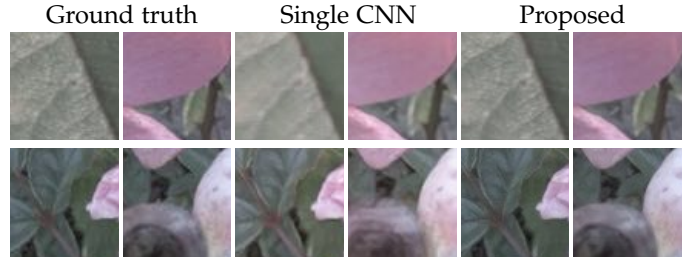


Fig. 2. Comparison against one single network. The single network has poorer textures and is unable to correctly model the geometry of the scene.

As it is reported in the table, the proposed loss function combining just the reconstruction error and gradients differences outperforms the other settings in both test datasets.

4.4 Comparison with One Single CNN

We compare the proposed approach against using one single CNN to model the view synthesis problem. The implemented single-CNN model consists of a fully-convolutional network of 22 layers with kernel sizes of 3×3 . Also, as in our disparity network, we use dilated convolutions from the fourth to the seventh layers at rates 2, 4, 8 and 16, respectively. This results in a network with 1.66 millions of parameters. The two images P and Q containing the angular coordinates and the four corner views are concatenated along the channels dimension and are the input to the network. The output is the color novel view at the indicated position. We have trained this model using the *Flowers* training set.

In Table 3 we quantitatively compare both models. The single CNN takes in average half a second less than the proposed approach but the performance is significantly worse. As seen in Figure 2, the single CNN reconstruction results are blurry and the network is unable to correctly model the geometry of the scene.

4.5 Advantage of Using Four Disparity Maps

Next, we show the importance of considering four different disparity estimations compared to the use of one common disparity map for the four views, as it is done in [14]. In Table 3 we quantitatively compare these two strategies. The use of multiple disparity estimations improves the estimation. In the *Flowers* dataset we obtain an average PSNR of 38.28, while with one common disparity it decreases to 37.76. On the other hand, in the *Diverse* test set we maintain a high PSNR of 38.12 with the proposed method, opposed to the 36.03 yielded by the use of one single disparity.

Figure 3 illustrates this comparison. The use of a common disparity leads to inaccuracies in the estimation that are

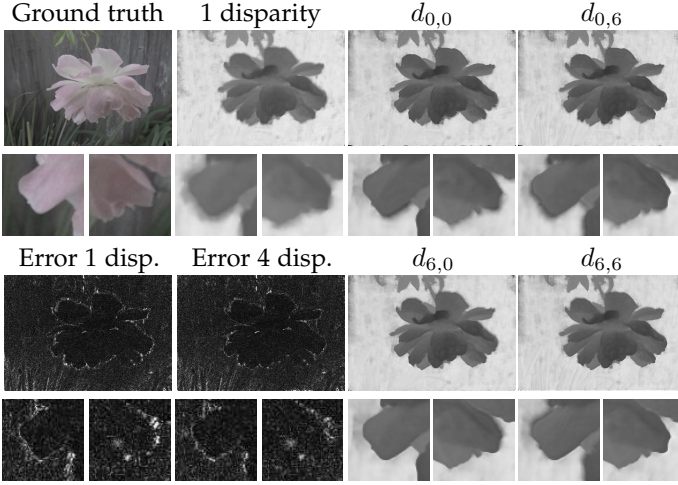


Fig. 3. Effect of using one common disparity map or the proposed multiple disparity maps. Reconstruction errors are clipped into the range $[0, 0.04]$ for images in $[0, 1]$. We encourage the reader to look at the electronic version of this paper to better see the details.

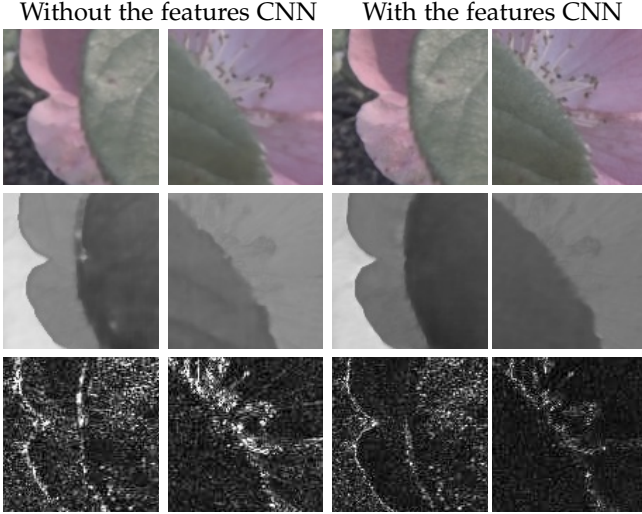


Fig. 4. Influence of the features network. We show the reconstructed center view (top), the disparity map $d_{0,0}$ (middle) and the error image (bottom), which is clipped into the range $[0, 0.04]$. We encourage the reader to look at the electronic version of this paper to better see the details.

located at the union of the occluded parts of the four images. That is, for instance, next to the boundaries of the flower. However, if we look to the case of having four disparities, we can see how the areas of the views corresponding to non-occluded regions are sharp and accurate, while the occluded parts present more difficulties. The effect in the final result is reflected in the error images, where the errors at occlusions are significantly smaller in the case of using four disparities.

4.6 Effect of Using the Features CNN

We now compare the proposed network against one that does not have a first stage for feature extraction and instead inputs to the disparity network are directly the light field views, as it is done for instance in [15]. In this case, in the disparity CNN we included more convolutional layers to have the same number of trainable variables. In Table 3 we can see

TABLE 4

Quantitative comparison with LBVS [14] and 4DLF [15]. The dataset in parenthesis after each method indicates the used training set, where F stands for *Flowers* and D for *Diverse*.

	<i>Flowers</i>			<i>Diverse</i>		
Method	MAE	PSNR	SSIM	MAE	PSNR	SSIM
LBVS (D)	1.374	34.37	0.9625	1.053	36.13	0.9799
4DLF (F)	2.998	33.10	0.9510	3.859	30.61	0.9369
Proposed (F)	0.878	38.29	0.9778	0.797	38.13	0.9849
Proposed (D)	0.982	37.34	0.9733	0.805	38.14	0.9846

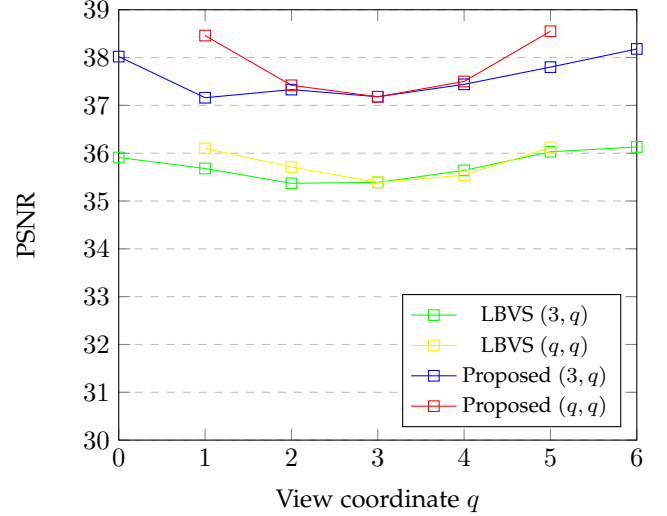


Fig. 5. Quality of the reconstruction for different views compared with LBVS [14]. The evaluation is in terms of the PSNR, which is averaged over the *Diverse* test set for each particular view. For each method, we show the results for the row of views in the middle, with coordinates $(3, q)$, and the evaluation for the views of the form (q, q) , that lie in the diagonal that goes from the view $(0, 0)$ to $(6, 6)$. Both methods have been trained on the same *Diverse* training set.

the gains in performance when using feature extraction, an average PSNR of 38.24 opposed to 37.81 when we do not use it in the *Flowers* test set. Moreover, on the *Diverse* test set it considerably decreases to 35.66. A visual example is shown in Figure 4. Without the features network the method produces inaccuracies in the disparity map, which results in a loss of textures and higher reconstruction errors.

4.7 Comparison against the State of the Art

Table 4 reports quantitative evaluation compared to the recent learning-based view synthesis from Kalantari et al. [14] (LBVS), that reconstructs any view of the light field given the four corner ones; and the approach proposed by Srinivasan et al. [15] (4DLF), that reconstructs de 4D light field given the center view. LBVS has been trained on the *Diverse* training set, while 4DLF was optimized on the *Flowers* one. The evaluation metrics reported in the table have been averaged over the indicated test set and over the intersection of the views that have to be estimated for the three methods. According to the table, the proposed approach outperforms the other methods in all the metrics and in both datasets.

In Figures 5 and 6 we plot the PSNR as a function of the view position (p, q) . In particular, we show the graph for the subset of views that are in the row of views in the

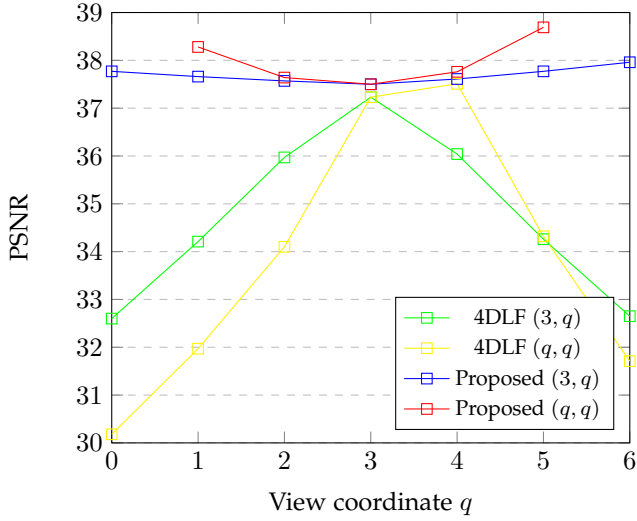


Fig. 6. Quality of the reconstruction for different views compared with 4DLF [15]. The evaluation is in terms of the PSNR, which is averaged over the *Flowers* test set for each particular view. For each method, we show the results for the row of views in the middle, with coordinates $(3, q)$, and the evaluation for the views of the form (q, q) , that lie in the diagonal that goes from the view $(0, 0)$ to $(6, 6)$. Both methods have been trained on the same *Flowers* training set.

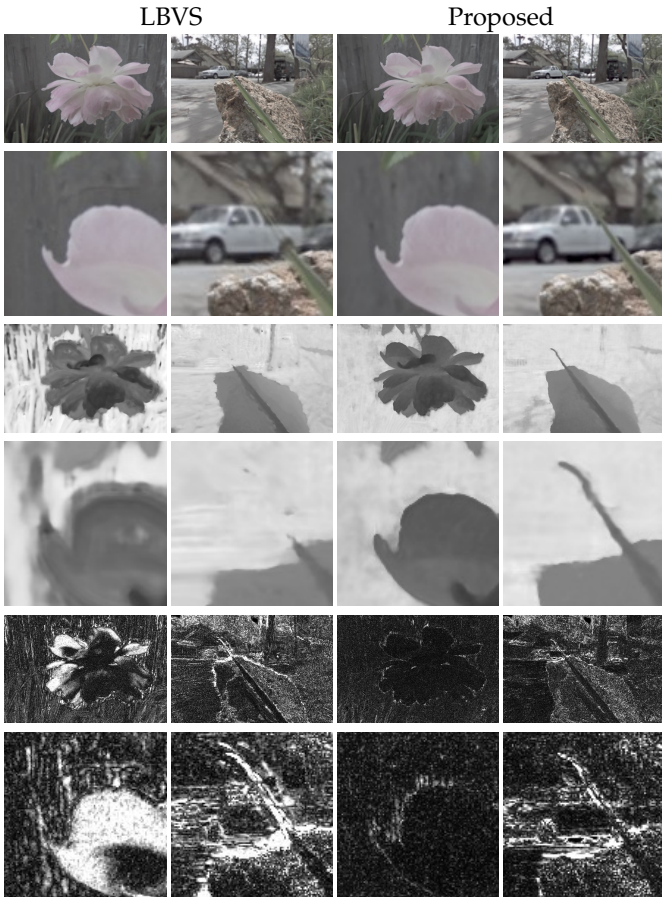


Fig. 7. Comparison against the state-of-the-art method LBVS [14] on two examples, one from the *Flowers* test set (left) and the other from *Diverse* (right). For each method, we show the synthesized view (top), the estimated disparity map (middle) and the reconstruction error (bottom), which is clipped into $[0, 0.04]$.

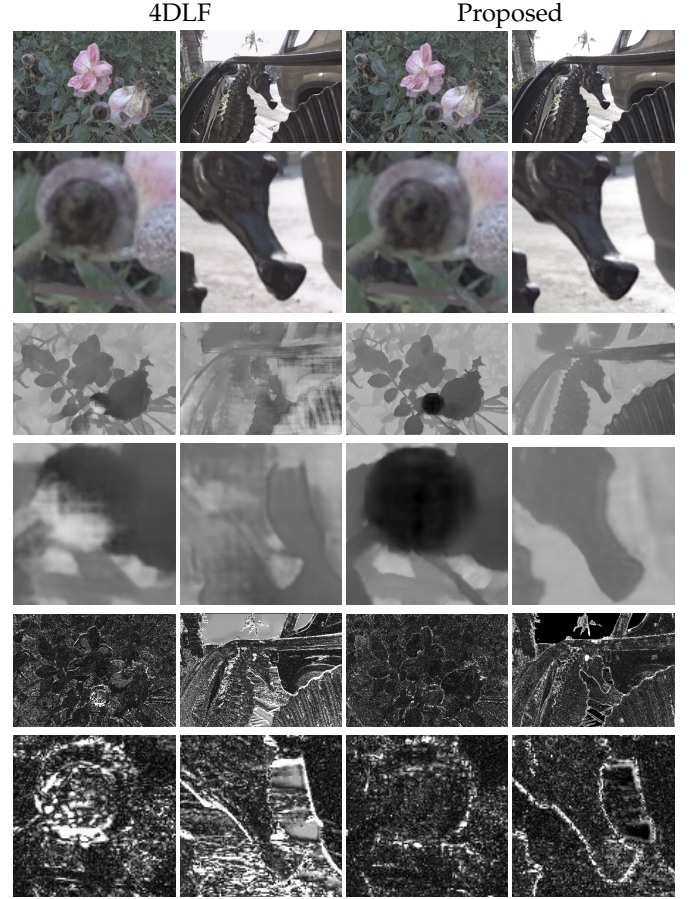


Fig. 8. Comparison against the state-of-the-art method 4DLF [15] on two examples, one from the *Flowers* test set (left) and the other from *Diverse* (right). For each method, we show the synthesized view (top), the estimated disparity map (middle) and the reconstruction error (bottom), which is clipped into $[0, 0.04]$.

middle of the light field, which have coordinates $(3, q)$, with $q \in \mathbb{Z} \cap [0, 6]$; and the graph for the views of the form (q, q) , with $q \in \mathbb{Z} \cap [0, 6]$, that lie in the diagonal that goes from the view $(0, 0)$ to $(6, 6)$. Figure 5 compares the proposed method with LBVS [14]. Both models have been trained on the *Diverse* training set and, for each view position, the PSNR values have been averaged over the *Diverse* test set. We can observe that, for both methods, a higher PSNR is reported for those views that are closer to the input ones. However, LBVS values differ from our metrics in more than one point for all views.

On the other hand, in Figure 6 we perform a similar comparison with 4DLF [15]. In this case, both models have been trained on the same *Flowers* training set and metrics are averaged over the *Flowers* test set for each angular position. According to the graph, the PSNR values of 4DLF drastically decrease as the novel view position distances from the input view, reaching values of almost 30. In our case, we also notice better results for those views that are closer to the input corner ones. However, differences in the PSNR values are smaller and always ranging between 37.50 and 38.70.

Figure 7 visually compares the result from LBVS [14] with ours. This method uses the same estimated disparity map to warp each corner view. Therefore, disparity maps are less accurate at depth discontinuities than in our case.

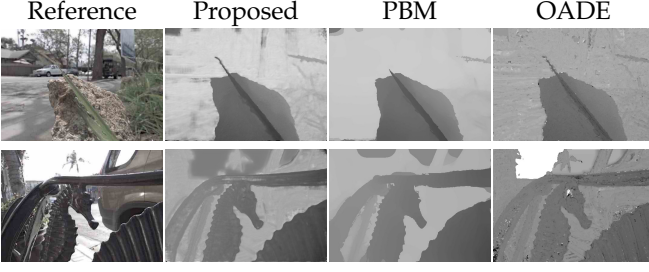


Fig. 9. Estimated disparity compared to state-of-the-art methods PBM [40] and OADE [41]. In our case, disparity is estimated from the four corner views, while methods PBM and OADE make use of the complete 4D light field.

As it can be seen in both crops, their reconstruction has difficulties at occluded regions, resulting in blurred results and artifacts at these areas. Moreover, the method in LBVS synthesizes the new view using a CNN that outputs the novel color image. In some cases this may produce changes in colors, as it can be noticed in the flower. Besides, their method is unable to recover the tip of the leaf since disparity is not correctly estimated in this thin structure.

In Figure 8 we compare our results with 4DLF [15]. Inferring a 4D light field from only one view may seem an advantage compared to our method. However, their method does not work properly with other images than flowers and it fails when dealing with complex scenes, where there is more than one object in the foreground. Although both methods have been trained on the same *Flowers* training set, their method is completely unable to model the geometry of the scene in some cases, resulting in high errors mostly located at object boundaries. On the contrary, our method better estimates disparity, which leads to smaller reconstruction errors.

4.8 Disparity Estimation

Although the purpose of this work is view synthesis, the proposed method can also be used to estimate disparity in an unsupervised manner. In Figure 9 we qualitative compare our estimated disparity to state-of-the-art methods for depth estimation from light fields. Specifically, the phase-based method from Jeon et al. [40] (PBM) and the occlusion-aware depth estimation from Wang et al. [41] (OADE). To estimate disparity, these methods make use of the complete 4D light field, while our approach only uses the four corner views. In spite of that, our method shows to be competitive with respect to both PBM and OADE.

4.9 Generalization

We assess the performance of the trained model tested on a different dataset by evaluating the model on the *Diverse* test set. Furthermore, we trained our networks from scratch on the *Diverse* training set and tested on both *Flowers* and *Diverse* test sets. Evaluation metrics for these experiments are reported on Table 4. Quantitatively, we can see how the method trained on *Flowers* yields a high PSNR also in the *Diverse* set. In addition, the training on the 100 images from the *Diverse* dataset yields similar performance than the one trained on *Flowers* on the testing images from the same

dataset, while just a slightly worse performance is observed in the *Flowers* test set.

4.10 Wide-Baseline Light Fields

Wide-baseline light fields make more difficult the problem of view synthesis than with Lytro images. Wider baselines involve having larger disparities and therefore much larger occluded areas. The proposed method specially treats the occlusion problem by assuming differences in the disparity maps and computing four different ones for each corner view. Moreover, we have seen that the selection network is able to detect occluded pixels and discard inaccurate reconstructions on these parts.

In this section, we apply the proposed approach to wide-baseline light fields. In particular, to light field images captured with the camera rig presented in [2]. The baseline between two consecutive cameras of this rig is 7cm. In this complex case we have to deal with larger disparities than with Lytro light fields. Therefore, we cannot directly use the same networks as the ones used in the previous case, since the receptive field of the disparity network will not be enough to match distant pixels. In the following we adapt the proposed networks to deal with this challenging case.

4.10.1 Adaptation to Wide-Baseline Light Fields

To increase the receptive field and provide the disparity CNN of more global information without introducing many parameters, we apply the same features CNN f_e as before at three different dilation rates for the first five convolutional layers. These dilations are 2, 4 and 8, respectively. The output volumes given from each dilation rate are concatenated along the depth dimension and fused by means of 1×1 convolutions to obtain a 32-channel feature volume.

The features CNN outputs a volume for each view $F_{0,0}, F_{0,N}, F_{N,0}$ and $F_{N,N}$. In the previous case, these features were concatenated and were the input to the disparity CNN. Here, as disparities and occluded areas are too large and trying to find correspondences between the four images at the same time might be too difficult for the network, we propose to compute disparity maps from horizontal and vertical pairs of views separately and then to fuse these disparity estimations by means of a simple convolutional network. With this strategy the disparity network can better establish matches between input images since the overlapping between horizontal or vertical pairs is larger than if we consider the four images at the same time.

Then, horizontal disparities are computed from the concatenation of horizontal pairs of views,

$$(d_{0,0}^h, d_{0,N}^h) = f_{d_h}(P, Q, F_{0,0}, F_{0,N}), \quad (14)$$

$$(d_{N,0}^h, d_{N,N}^h) = f_{d_h}(P, Q, F_{N,0}, F_{N,N}); \quad (15)$$

while vertical ones are computed from vertical pairs,

$$(d_{0,0}^v, d_{N,0}^v) = f_{d_v}(P, Q, F_{0,0}, F_{N,0}), \quad (16)$$

$$(d_{0,N}^v, d_{N,N}^v) = f_{d_v}(P, Q, F_{0,N}, F_{N,N}). \quad (17)$$

Functions f_{d_h} and f_{d_v} are convolutional networks that have the same architecture as the disparity CNN (f_d) but replacing input and output sizes with 64 and 2 channels, respectively.

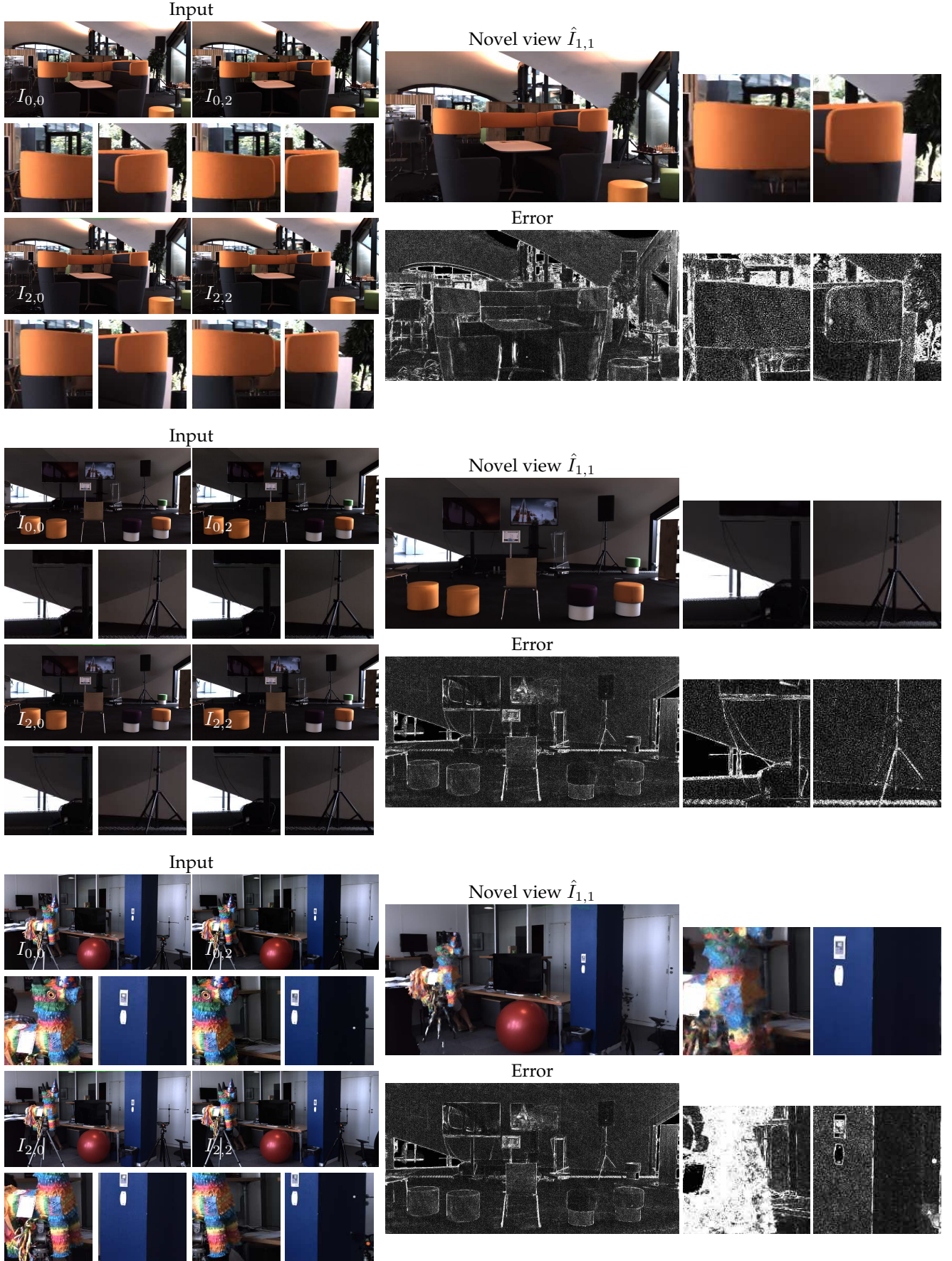


Fig. 10. Visual results on light fields acquired with the camera rig presented in [2]. Error images are clipped into the range $[0, 0.04]$ for images in $[0, 1]$.

Disparities estimated from horizontal and vertical displacements are fused into a single disparity map for each view,

$$d_{s,t} = f_{df}(d_{s,t}^h, d_{s,t}^v), \quad s, t \in \{0, N\} \quad (18)$$

with f_{df} being a convolutional neural network of two layers with kernels 3×3 and 1×1 , respectively.

Once we have computed the four disparity maps, the algorithm follows as before. The four views are warped using the corresponding disparity according to Equation (5). Then, images P and Q , the warped views $I_{s,t}^w$, with $s, t \in \{0, N\}$ and disparity maps $d_{s,t}$, with $s, t \in \{0, N\}$ are concatenated along the channel dimension and are the input to the selection network f_s . The selection network is exactly the same as in the Lytro case. Finally, the predicted center view is computed as a weighted average of the four warped views using as weights the selection masks, according to Equation (8).

The receptive field of the network proposed for the wide-baseline case is 170 pixels, compared to 97 pixels for the plenoptic version. The full model has a total of 2.02 million of parameters to learn. As training loss function, we use the one from Equation (12) with two additional terms that enforce the warped views with the horizontal and vertical disparities to be similar to the ground truth image.

4.10.2 Training details

We train these networks from scratch on light field images captured with the camera rig presented in [2]. Video sequences from indoor and outdoor scenarios have been recorded and one of every ten frames has been selected as training light field. These light fields have been rectified and viewpoints are arranged on a regular grid. From the available 4×4 views, we randomly select an array of 3×3 and, from the four corner images, we train the networks to estimate the center one. Also, these light fields have a spatial resolution of 2048×1088 and are spatially downsampled by a factor of 2.

The training set contains 212 light fields and, considering that we take subsets of 3×3 views, this results in a total of 848 examples. We randomly extract patches of 250×250 from the training images to train the model. The network is optimized using the ADAM solver [37] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, a learning rate of 0.001 and a batch size of 1. Weights are randomly initialized using the Xavier method [38] and the softmax β is initialized to 1. The maximum disparity has been set to $d_{\max} = 60$. The method converges after 300 thousand iterations, which approximately takes 2 days and 20 hours on a GeForce GTX 1080 Ti GPU. At test time, it takes 20 seconds to synthesize a new instance.

4.10.3 Application to Wide-Baseline Light Fields

Figure 10 illustrates an example of view synthesis for different light fields captured with the camera rig presented in [2], using the proposed method. The method shows very promising results as it can be seen in the figure. The crops from the input views give an intuition of how large are occlusions in each case. By looking at the error images, in general most of these occluded parts do not present large errors. In the first and second examples, largest errors are

present in some parts of the background, mainly on bright areas.

In the last example, we have two objects with large disparities and the method is unable to correctly estimate them. This results in a blurred reconstruction and a thin structure that does not appear in the predicted center view. This suggests that the receptive field of the network is not enough to deal with these large disparities. However, the method that was first designed to cope with plenoptic cameras generally yields promising results for this challenging case, being able to detect from each view that parts that are visible in the center one.

5 CONCLUSIONS

In this work, we proposed a novel learning-based approach for new view synthesis for light field images. In particular, given the four corner views of a light field, we have tackled the problem of estimating any view in between. The method uses three sequential networks for feature extraction, for disparity estimation and another for view selection. Compared to the state of the art, we propose to compute four different disparity maps in order to deal with the occlusions problem. Experiments have demonstrated the importance of using this strategy, jointly with the selection network, to obtain accurate results at occlusions. The method has proved to outperform the state of the art for Lytro light fields and its application to light fields from the camera rig from [2] has given very promising results.

As future work, we plan to focus on wide-baseline light fields and work on architectures for this special case, in which networks should incorporate more context information in order to deal with large disparities and occlusions.

ACKNOWLEDGMENTS

J. Navarro acknowledges support from Ministerio de Economía y Competitividad of the Spanish Government under grant TIN2017-85572-P (MINECO/AEI/FEDER, UE).

REFERENCES

- [1] R. Ng, "Digital light field photography," Ph.D. dissertation, Stanford, CA, USA, 2006, aAI3219345, Thesis led to commercial light field camera (Lytro camera). [Online]. Available: www.lytro.com
- [2] N. Sabater, G. Boisson, B. Vandame, P. Kerbiriou, F. Babon, M. Hog, R. Gendrot, T. Langlois, O. Bureller, A. Schubert *et al.*, "Dataset and pipeline for multi-view lightfield video," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 1743–1753.
- [3] L. Dabala, M. Ziegler, P. Didyk, F. Zilly, J. Keinert, K. Myszkowski, H.-P. Seidel, P. Rokita, and T. Ritschel, "Efficient multi-image correspondences for on-line light field video processing," in *Computer Graphics Forum*, vol. 35, no. 7. Wiley Online Library, 2016, pp. 401–410.
- [4] R. S. Overbeck, D. Erickson, D. Evangelakos, M. Pharr, and P. Debevec, "A system for acquiring, processing, and rendering panoramic light field stills for virtual reality," *arXiv preprint arXiv:1810.08860*, 2018.
- [5] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," *arXiv preprint arXiv:1703.04309*, 2017.
- [6] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418.

- [7] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2758–2766.
- [8] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with bm3d?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2392–2399.
- [9] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 184–199.
- [10] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 24–32.
- [11] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4748–4757.
- [12] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldluecke, "Light field intrinsics with a deep encoder-decoder network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 28, p. 15, 2018.
- [13] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deepstereo: Learning to predict new views from the world's imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5515–5524.
- [14] 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.
- [15] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, "Learning to synthesize a 4d rgb-d light field from a single image," in *Proceedings of the International Conference on Computer Vision (ICCV)*, vol. 2, 2017, p. 6.
- [16] 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, p. 30, 2013.
- [17] M. Goesele, J. Ackermann, S. Fuhrmann, C. Haubold, R. Klawnsky, D. Steedly, and R. Szeliski, "Ambient point clouds for view interpolation," in *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, ACM, 2010, p. 95.
- [18] A. Fitzgibbon, Y. Wexler, and A. Zisserman, "Image-based rendering using image-based priors," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 141–151, 2005.
- [19] E. Shechtman, A. Rav-Acha, M. Irani, and S. Seitz, "Regenerative morphing," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 615–622.
- [20] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606–619, 2014.
- [21] 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.
- [22] Z. Zhang, Y. Liu, and Q. Dai, "Light field from micro-baseline image pair," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3800–3809.
- [23] E. Penner and L. Zhang, "Soft 3D reconstruction for view synthesis," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, p. 235, 2017.
- [24] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on epi," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2017, 2017, p. 2.
- [25] 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)*, September 2018.
- [26] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," *arXiv preprint arXiv:1805.09817*, 2018.
- [27] P. P. Srinivasan, R. Tucker, J. T. Barron, R. Ramamoorthi, R. Ng, and N. Snavely, "Pushing the boundaries of view extrapolation with multiplane images," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [28] 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 (SIGGRAPH)*, 2019.
- [29] Z. Liu, R. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proceedings of the International Conference on Computer Vision (ICCV)*, vol. 2, 2017.
- [30] J. van Amersfoort, W. Shi, A. Acosta, F. Massa, J. Totz, Z. Wang, and J. Caballero, "Frame interpolation with multi-scale deep loss functions and generative adversarial networks," *arXiv preprint arXiv:1711.06045*, 2017.
- [31] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," *arXiv preprint arXiv:1803.10967*, 2018.
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [33] D. Fourure, R. Emonet, E. Fromont, D. Muselet, A. Tremeau, and C. Wolf, "Residual conv-deconv grid network for semantic segmentation," *arXiv preprint arXiv:1707.07958*, 2017.
- [34] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super slomo: High quality estimation of multiple intermediate frames for video interpolation," *arXiv preprint arXiv:1712.00080*, 2017.
- [35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
- [36] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, vol. 16, 2016, pp. 265–283.
- [37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [38] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.
- [39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [40] 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*, 2015, pp. 1547–1555.
- [41] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3487–3495.

Julia Navarro received the BSc in Mathematics from the Universitat de les Illes Balears, Spain, in 2014 and the MSc in Computer Vision from the Universitat Autònoma de Barcelona, Spain, in 2015. She is currently pursuing the PhD degree with the Universitat de les Illes Balears. Her research interests are computer vision and mathematical image analysis.

Neus Sabater received the BSc degree in 2005 from the Universitat de Barcelona, Spain, and the MSc and PhD degrees in 2006 and 2009, respectively, in image processing from the Ecole Normale Supérieure de Cachan, France. She was a postdoctoral researcher at the California Institute of Technology before being appointed at Technicolor Research & Innovation in 2011 where she is currently a Senior Scientist. Her research interest include image processing, computer vision and computational photography.