Deep Light Field Acquisition Using Learned Coded Mask Distributions for Color Filter Array Sensors

Guillaume Le Guludec, Ehsan Miandji, Christine Guillemot

Deep Light Field Acquisition Using Learned Coded Mask Distributions for Color Filter Array Sensors

Guillaume Le Guludec, Ehsan Miandji, Christine Guillemot, *Fellow, IEEE*

***Abstract*—Compressive light field photography enables light field acquisition using a single sensor by utilizing a color coded mask. This approach is very cost effective since consumer-level digital cameras can be turned into light field cameras by simply placing a coded mask between the sensor and the aperture plane. This paper describes a deep learning architecture for compressive light field acquisition using a color coded mask and a sensor with Color Filter Array (CFA). Unlike previous methods where a fixed mask pattern is used, our deep network learns the optimal distribution of the color coded mask pixels. The proposed solution enables end-to-end learning of the color-coded mask distribution and the reconstruction network, taking into account the sensor CFA. Consequently, the resulting network can efficiently perform joint demosaicing and light field reconstruction of images acquired with color-coded mask and a CFA sensor. Compared to previous methods based on deep learning with monochrome sensors, as well as traditional compressive sensing approaches using CFA sensors, we obtain superior color reconstruction of the light fields.**

***Index Terms*—Light Field imaging, compressed sensing, deep learning, inverse problems**

I. Introduction

Light field imaging has recently gained interest in the research community, due to its potential for a variety of applications, going from computational photography, e.g. by enabling genuine post-capture refocusing, to medical imaging and virtual and augmented reality. While in conventional 2D imaging, each sensor element sums all the light rays emitted by one 3D scene point over the lens aperture, *i.e.* records a 2D projection of the 3D points on the image plane, light fields instead record the radiance along each ray emitted by the 3D points according to different orientations. The light field can hence be seen as capturing an array of viewpoints (or sub-aperture images) of the scene, leading to a 4D ray-based scene representation. However, capturing a light field in a way that is memory and computationally efficient, as well as accurate, is challenging. Several camera architectures have been designed to capture light fields. Early attempts at capturing high resolution light fields include large camera arrays [41] or a single camera placed on a moving gantry [22]. While these devices can produce high quality images both in the spatial and angular domain, they are in general quite bulky and costly in terms of storage. More practical solutions include light-weight devices like lenslet-based cameras [30] or angle sensitive pixel cameras [16]. These devices, however, sacrifice spatial resolution for angular resolution, hence capture views with a significantly lower spatial resolution compared to traditional 2D cameras.

More recent light field camera designs consider coded masks instead of micro-lens arrays to modulate 4D light fields into 2D projections that can be captured by a digital camera sensor. Reconstruction algorithms are used, based on the compressed sensing paradigm [7], to restore the original light field from its projections [4, 25, 27]. Compressed sensing provides a theoretical framework that enables the reconstruction of high-dimensional data from lower-dimensional projections when assuming additional signal constraints such as sparsity in a particular transform domain. Coded mask acquisition techniques are cost effective since a consumer-level digital camera can be used for light field acquisition by placing a coded mask in front of the sensor.

While compressed sensing originally provides theoretical guarantees for the signal reconstruction under signal sparsity assumptions [28, 9, 36], it has been empirically shown that deep priors on the signal, like the deep generative hypothesis [6], outperform the sparsity priors for a great number of image reconstruction tasks. The authors of [6] also provide theoretical evidence that deep priors can replace the traditional sparsity hypothesis. Moreover, reconstruction techniques based on deep networks usually allow for a reconstruction of the signal in one forward pass through a deep model, making them orders of magnitude faster than traditional iterative methods like the orthogonal matching pursuit [32].

Deep learning architectures have been proposed in [37, 12, 29] for light field reconstruction from a sparse set of measurements recorded on a monochrome sensor. For compressive light field acquisition, Nabati et al. [29] were able to get state-of-the-art results with a monochrome sensor using a fully convolutional network and a low-entropy random red-green-blue-white (RGBW) color-coded mask. In [11], the authors propose a convolutional network architecture to compute the coded sub-aperture images. The light field reconstruction problem is then solved using an iterative optimization approach with a deep spatio-angular regularization prior. The end-to-end pipeline is applied to each color channel independently. None of the previous work takes into account the fact that camera sensors are usually equipped with a color filter array (CFA) [5] performing a form of color compression. Much research has been done in trying to find optimal CFAs and demosaicing algorithms for 2D imaging. The authors of [15]

The paper has been submitted for review on Dec. 1st. This work was supported in part by the EU H2020 Research and Innovation Programme under grant agreement No 694122 (ERC advanced grant CLIM), and in part by the french ANR research agency in the context of the artificial intelligence project DeepCIM.

The authors are with Inria Rennes – Bretagne-Atlantique
263 Avenue Général Leclerc, 35042 Rennes Cedex, France (e-mail: first-name.lastname@inria.fr).

obtained state-of-the-art results on traditional 2D images by jointly learning a deep demosaicing network and optimizing the CFA itself. Note that the learning of optimal masks has also been considered in [8], [34], [44], but for completely different problems. The authors in [8] address the problem of jointly optimizing a color filter array for 2D image sensors and of the de-mosaicing process. The authors in [34] aim at extending the depth of field from 2D captures by proposing a wave-based image formation model and a joint optimization of a diffractive optical element together with a de-convolution method. The problem addressed in [44] is monocular scene depth estimation by using learned phase masks. None of the three methods addresses the problem of light field acquisition with optimized color coded masks that we consider here.

In this paper, we propose a deep learning architecture for an end-to-end optimization of both the color-coded mask distribution and the reconstruction algorithm. The problem of optimizing the coded mask has previously been addressed in [25] by minimizing the mutual coherence of the equivalent sensing matrix, and in [11] by using deep learning techniques. In contrast to this prior work, we propose a new approach based on learning the distribution of the mask pixels. Indeed, the learned distribution admits a higher representation power compared to a fixed mask pattern learned over a given training set. In other words, the mask distribution generalizes well to unseen test data, as shown by our results in Section V. Moreover, it has been shown that adaptive sampling, e.g. via learning a mask pattern, cannot outperform naive random sampling in terms of Mean Square Error (MSE) of the estimation [1]. In addition, unlike previous methods where the mask is learned independently, our proposed framework jointly optimizes the mask distribution and the reconstruction network in an end-to-end optimization approach.

We also consider the case where the sensor is equipped with a color filter array which is taken into account in the optimization of the color coded mask distribution. We show that the proposed framework can efficiently perform joint demosaicing and light field reconstruction of images acquired with a color-coded mask and a CFA-equipped sensor. Experimental results show that using a CFA, compared to monochrome sensors as in [29], enables a superior reconstruction quality of a wide variety of light field images.

In summary, our contributions are as follows:

- We propose a novel reconstruction architecture for light field acquisition with learned color coded masks. We show that the results obtained with the proposed architecture are superior, with an average PSNR gain of 1.36 dB, when compared to state-of-the-art methods.
- We then propose an end-to-end learned generator of the color coded mask distribution, which is shown to further improve the reconstruction quality.
- We also introduce a Bayer CFA in the acquisition and reconstruction pipeline and show that, when combined with the learned color coded mask, it further improves the reconstruction quality. The complete pipeline outperforms by almost 2 dB the state-of-the-art method.
- We further include the learning of the color filter array in the pipeline and show that it improves over using a Bayer CFA, especially in presence of noise.
- Finally, we show that the approach is robust to noise when the noise level is low. However, when the noise level increases, the learned CCM converges towards a uniform transparent mask, impacting the overall reconstruction performance. To cope with this problem, we further introduce an entropy-based regularization of the coded-mask and show that this regularization constraint leads to an average of 2 dB improvement in presence of a high level of noise.

II. Related Work

Mask-based cameras with compressed sensing

Programmable aperture approaches have first been considered to sequentially capture subsets of light rays. The idea consists in time-multiplexing 2D slices of the 4D light field on the sensor, using a programmable non-refractive mask placed at the aperture as in [23]. The latter design exploits the fast multiple-exposure feature of digital sensors. The authors in [38] use instead optical heterodyning to frequency multiplex the 4D Fourier transform of the light field into spatio-angular bands on the 2D sensor. A good overview of the above camera designs can be found in [40].

Since the light field data is typically high dimensional and compressible, its acquisition can be placed in a compressive sensing framework, in which the sensing matrix is materialized by a coded physical mask. Thanks to the use of a coded mask, instead of recording a spatial multiplex of 2D slices of the light field, as in micro-lens array based camera architectures, the photosensor records a set of linear measurements from which a higher resolution light field can be reconstructed. This compressive sensing principle is applied in [43, 42] where the 2D sensor captures optically coded projections using two attenuation masks separately placed at the aperture plane and in front of the sensor. Given the measurements recorded on the sensor, the light field is then reconstructed using a least square minimization with a total variation regularization constraint. Similarly, the authors in [3] place a randomly coded mask on the aperture plane to obtain incoherent measurements of the light field. Multiple shots are captured as random linear combinations of angular images by separately opening one region of the aperture and blocking the light in the others.

The authors in [25] propose a camera architecture that records optically coded projections on a single image sensor using a monochrome mask, while the authors in [26] and [27] use respectively a random stationary or a moving color-coded mask to extract incoherent measurements. In both cases, the light field is then reconstructed using a compressive sensing framework, assuming that the light field is sparse in a domain defined by an overcomplete dictionary [25], [27] or an ensemble of 2D separable dictionaries [26]. The authors in [31] introduce an Equivalent Multi-Mask Camera (EMMC) model which unifies most existing single lens mask-based light field cameras, allowing for a flexible configuration of a variety of sensing schemes.

Mask-based cameras with deep reconstruction

While the first solutions were considering classical sparse reconstruction methods, the problem of light field reconstruction can also be efficiently solved using deep learning techniques [37, 12, 29, 17]. The authors in [37, 12, 29] assume a pre-defined and fixed mask pattern and propose convolutional neural network architectures to reconstruct the light field from the set of measurements using the coded mask. Given coded measurements, the method in [12] generates two coarse light fields which are then fused to generate the final estimate of the original light field.

The authors in [29] introduce a sensing matrix which modulates both color and angular information of a light field into 2D sensor measurements. The sensing matrix together with the coded measurements are fed into a CNN-based network to reconstruct the light field. In contrast, the authors in [17] pose the coded aperture acquisition and light field reconstruction as an auto-encoder and optimize the mask pattern together with the parameters of the reconstruction algorithm in an end-to-end auto-encoder learning. A convolutional kernel of size 1×1 is used to simulate the coded aperture process. Then, two sequential sub-networks, the second one being based on the VDSR network [19], are used to reconstruct the light field from the coded measurements. Our acquisition model in this paper is distinct from [17] since we place a mask between the aperture plane and the sensor. A learned convolutional network architecture is used in [11] to compute the coded sub-aperture images, from which the light field is reconstructed using an iterative optimization approach with a deep spatio-angular regularization prior. The end-to-end pipeline is applied on each color channel separately.

Unlike the above solutions that assume the sensor to be monochrome, use a fixed mask pattern, or independently optimize a coded aperture pattern, here we propose an end-to-end learning framework to jointly optimize the color-coded mask distribution and the reconstruction network. We also consider the case where the sensor is equipped with a CFA, which is commonly used for digital consumer-level cameras.

Phase-mask coded aperture camera designs

The authors in [10] introduce a solution for enhanced depth-of-field based on a binary phase-mask composed of a ring pattern, whereby each ring introduces a different phase-shift to the wavefront emerging from the scene. The phase-mask is designed such that an end-to-end trained CNN is able to restore from the aperture coded image an all-in-focus image, hence with a large depth of field. A layer of the CNN models the phase-mask parameters (ring radii and phase). A similar phase-coded aperture camera is proposed in [13] for monocular depth estimation. An optical phase mask provides depth-related color characteristics that are then used for estimating scene depth with a fully convolutional neural network. Given that depth-dependent defocus 'bokeh', or the point spread function, depends on the amplitude and phase of the aperture, a solution is proposed in [44] to end-to-end optimize a phase mask and an algorithm that allows accurate scene depth estimation from a single viewpoint. Light propagation from the scene to the sensor and the modulation by the mask are modeled as front-end layers of a deep neural network.

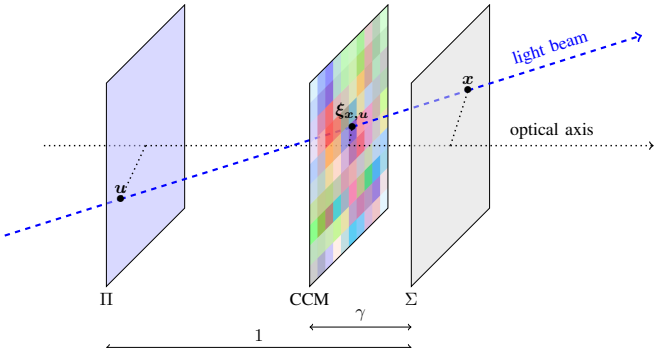


Fig. 1: **Two-plane parameterization with color-coded mask.** A light beam is characterized by the coordinates of its intersection with the reference plane Π and the sensor plane Σ, denoted respectively $\boldsymbol{u}$ and $\boldsymbol{x}$. The beam intersects the mask plane at coordinates $\boldsymbol{\xi}_{\boldsymbol{x},\boldsymbol{u}} = (1-\gamma)\boldsymbol{x} + \gamma\boldsymbol{u}$.

The problem addressed in [8] is instead the design of color filter arrays for 2D image sensors, by end-to-end learning of the color pattern and of the image reconstruction which is therefore a de-mosaicing problem. The authors in [34] propose a wave-based image formation model for 2D cameras and an approach for joint optimization of a diffractive or refractive element together with a de-convolution method for achromatic extended depth of field and snapshot super-resolution 2D imaging.

In contrast, the proposed approach differs from all the above methods, since it concerns compressive light field acquisition, and does not concern coded aperture cameras using phase masks nor problems of depth estimation or depth of field extension with 2D cameras.

III. Mathematical framework

A. Light fields

We adopt the two-plane parameterization of light fields, in which any light field can be represented by a collection of light rays passing through two points on a pair of parallel planes. In addition to a sensor plane Σ, we consider a reference plane Π that usually coincides with the aperture plane. A light ray is then defined by a set of coordinates on the sensor plane (x, y), called the *spatial coordinates*, together with a set of coordinates on the reference plane (u, v) called the *angular coordinates*, see Figure 1. The first set of coordinates serves to describe the point of incidence of the ray on the sensor plane Σ, whereas the second set (u, v) defines the direction of the light ray. To simplify the notations, in the remainder of the paper we use $\boldsymbol{x} = (x, y)$ to designate the pair of spatial coordinates, and $\boldsymbol{u} = (u, v)$ to designate the pair of angular coordinates.

In this framework, a light field is a real function of a five-dimensional real vector space defined as

$$\begin{aligned} \mathcal{L}\colon \Sigma \times \Pi \times \Lambda \simeq \mathbf{R}^5 &\to \mathbf{R} \\ (\boldsymbol{x}, \boldsymbol{u}, \lambda) &\mapsto \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}, \lambda), \end{aligned}$$

where λ corresponds to a given wavelength and Λ is the space of all possible wavelengths. The projection of the light field on the sensor plane is then given by

$$\mathcal{I}(\boldsymbol{x}) = \int_{\Pi\times\Lambda} \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}, \lambda)\phi(\lambda)\mathrm{d}\boldsymbol{u}\mathrm{d}\lambda, \tag{1}$$

where ϕ weights the sensitivity of the sensor to various wavelengths. For simplicity, we also consider the vignetting effect to be included in $\mathcal{L}$.

B. Compressed light field acquisition

Compressed light field acquisition aims at reconstructing a light field using lower-dimensional projections captured on the sensor plane. The compressed sensing theory indeed tells us that, under conditions of incoherence between the measurement space and the signal space, a signal can be recovered from a sparse set of measurements, provided the signal is sparse in a particular transform domain. The incoherence property is satisfied by taking random measurements. Concerning light field acquisition, this implies applying an additional linear modulation of the light field before the projection. This can be expressed as

$$\mathcal{I}(\boldsymbol{x}) = \int_{\Pi\times\Lambda} \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}, \lambda)M(\boldsymbol{x}, \boldsymbol{u}, \lambda)\mathrm{d}\boldsymbol{u}\mathrm{d}\lambda \tag{2}$$

where M represents a modulation of the light field that may depend on the spatial coordinates, the angular coordinates, and the wavelength.

1) Color-coded mask: We are interested in the case where the modulation is performed by a *color-coded mask* (CCM) placed in front of the sensor plane. The mask performs some filtering or light attenuation and can be defined by a function $m_{\mathrm{CCM}} : (u, \lambda) \mapsto m_{\mathrm{CCM}}(u, \lambda)$. Simple geometric considerations show that in this case

$$\mathcal{I}(\boldsymbol{x}) = \int_{\Pi\times\Lambda} \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}, \lambda) \cdot m_{\mathrm{CCM}}(\boldsymbol{\xi}_{\boldsymbol{x},\boldsymbol{u}}, \lambda)\mathrm{d}\boldsymbol{u}\mathrm{d}\lambda \tag{3}$$

where we assume the distance between the reference and sensor planes to be unit, and $\xi_{\boldsymbol{x},\boldsymbol{u}} = (1-\gamma)\boldsymbol{x} + \gamma\boldsymbol{u}$ with γ being the distance between the sensor and the mask. Note that this is equivalent to the previous equation with the additional constraint that the modulating function satisfies $M(\boldsymbol{x}, \boldsymbol{u} + \Delta\boldsymbol{u}, \lambda) = M(\boldsymbol{x} + \gamma\Delta\boldsymbol{u}, \boldsymbol{u}, \lambda)$. Figure 1 gives a visual explanation for equation 3.

2) Color filter array: The above equations implicitly assume monochromatic sensors. Nonetheless, it is possible to rewrite these formulae such that the acquisition is performed using a sensor equipped with a color filter array. A *color filter array* is a grid placed directly on the sensor plane that defines how much of every wavelength a given pixel will be sensitive to. It can thus be defined, much like the CCM, by a function $m_{\mathrm{CFA}} : (\boldsymbol{x}, \lambda) \mapsto m_{\mathrm{CFA}}(\boldsymbol{x}, \lambda)$. CFAs are generally periodic, and the most widespread CFA pattern is probably the Bayer pattern.

3) Effective modulating function: In the presence of both a color-coded mask m_{CCM} and a CFA m_{CFA}, the sensing equation becomes

$$\mathcal{I}(\boldsymbol{x}) = \int_{\Pi\times\Lambda} \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}, \lambda)m_{\mathrm{CCM}}(\boldsymbol{\xi}_{\boldsymbol{x},\boldsymbol{u}}, \lambda)m_{\mathrm{CFA}}(\boldsymbol{x}, \lambda)\mathrm{d}\boldsymbol{u}\mathrm{d}\lambda \tag{4}$$

$$= \int_{\Pi\times\Lambda} \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}, \lambda)M_{\mathrm{effective}}(\boldsymbol{x}, \boldsymbol{u}, \lambda)\mathrm{d}\boldsymbol{u}\mathrm{d}\lambda, \tag{5}$$

where M is the effective mask resulting from the product of both the CFA that performs color compression and the CCM that performs the angular compression, as well as color compression. Because the actual measurement of the image is performed on a finite grid, the above equation can be re-written in a discretized form as

$$\mathcal{I}(\boldsymbol{x}) = \sum_{\boldsymbol{u},\lambda} \mathcal{L}(\boldsymbol{x}, \boldsymbol{u}, \lambda) \cdot M_{\mathrm{effective}}(\boldsymbol{x}, \boldsymbol{u}, \lambda), \tag{6}$$

with the sum being over a finite set of size $\omega \cdot \nu \cdot \kappa$, where ω is the overall spatial resolution, ν is the angular resolution, and κ is the spectral resolution, that is, the number of color channels. In what follows, we assume $\kappa = 3$.

IV. Deep acquisition and reconstruction architecture

We describe an architecture that allows us to define and train a deep convolutional network to reconstruct a light field from compressed measurements recorded by a CFA-equipped sensor, along with the optimization of the CCM distribution. The end-to-end architecture for mask generation, as well as the compressed acquisition and reconstruction, is depicted in Figure 2.

We consider a (discretized) light field as a 5D tensor of dimensions $(\omega_x, \omega_y, \nu_u, \nu_v, \kappa)$. Similarly, the discretized version of the effective modulating function M can be considered as a 5D tensor of same dimensions. Note that the values of the effective modulating function correspond to the values of a uniform light field modulated by this function. The sensing equation in (6) can be functionally reformulated as

$$\mathcal{I} = \mathsf{Sensing}(\mathcal{L}, M), \tag{7}$$

where $\mathsf{Sensing}(.)$ is a bilinear operator performing the element-wise multiplication of the tensors followed by summation over the angular and spectral domains. Note that in this formalism, the measurement matrix is implicitly defined, since $\mathsf{Sensing}(.)$ is a bilinear operator. The authors of [29] use this formalism to devise a simple reconstruction scheme. The idea is to simply feed the sensed image along with the effective modulating function values in a way that is readily processable by a feed-forward fully convolutional network. The values of the modulating function are presented as a $(\omega_x, \omega_y, \nu \cdot \kappa)$ tensor, which is interpreted as a collection of feature maps, each feature map corresponding to a color channel of a sub-aperture view. The modulating values are then concatenated with the sensed image to yield a $\nu \cdot \kappa + 1$-channels image. Our idea is to learn the values of the modulating function together with the light field reconstruction network.

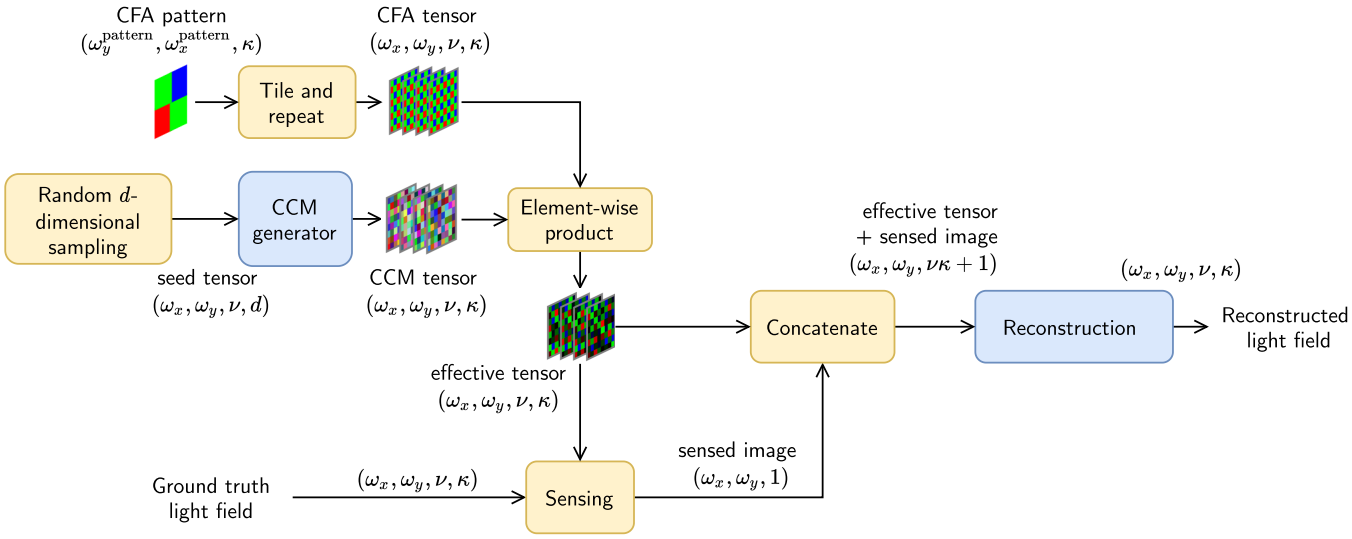


Fig. 2: **End-to-end mask generation, acquisition and reconstruction architecture.** The color-coded mask generator is first fed a $(\omega_x, \omega_y, \nu, d)$ tensor, with d being the dimension of an individual seed. It then outputs a $(\omega_x, \omega_y, \nu, \kappa)$ tensor corresponding to the CCM modulating function, where $\kappa = 3$ is the number of color channels. The tensor corresponding to the modulating function of the CFA is simply the repetition of the 2D CFA pattern for every view (sub-aperture images). The tensor corresponding to the effective modulating function is then computed by taking the element-wise product of the two tensors. The sensed image is subsequently simulated by the Sensing operator using the effective tensor and the input light field. Finally the effective sensing tensor is concatenated to the sensed image for producing a $(\omega_x, \omega_y, \nu\kappa + 1)$ tensor to be fed into the fully convolutional reconstruction network.

A. *Modulating tensor generation*

We consider color-coded masks in which the transmittance of all pixels are drawn from the same distribution and are independent from one another. The modulation is processed by element-wise multiplication with a stack of sub-aperture modulating images which are translated versions of each others, with a translation proportional to γ. Therefore, only pixels sufficiently far apart will be correlated, which ensures that the values of the modulation function are locally independent and identically distributed. Given that the CNN will not be able to capture correlation between distant pixels, we can assume that the elements of the modulating functions corresponding to a CCM are i.i.d. without sacrificing the accuracy.

We therefore seek to find a suitable probability distribution for the transmittance of the modulating tensor pixels. To this end, we train a feed-forward generator network

$$\text{seed} \mapsto f_{\text{gen}}(\text{seed}, \theta) \tag{8}$$

to output a κ-dimensional transmittance for each pixel, that is, an element of $[0, 1]^\kappa$, given a random seed drawn from a distribution in a d-dimensional space (in our experiments, we used the standard normal distribution $\mathcal{N}(0, 1)$ on $\mathbf{R}^d$). This function is then applied element-wise on a random $(\omega_x, \omega_y, \nu_u, \nu_v)$ tensor to yield the modulating tensor $(\omega_x, \omega_y, \nu_u, \nu_v, \kappa)$. The effective modulating tensor is then computed as the entry-wise product of the previously generated color-coded mask tensor with the modulating tensor of the CFA, which consists of identical copies of the color array for each sub-aperture view, see Figure 2. The process of generating the effective modulating tensor is differentiable, and thus, when linked to a subsequent differentiable reconstruction network, allows for an end-to-end implementation and learning of the mask generation and light field reconstruction pipeline. In all our experiments, we used a dense generator network composed of 3 dense hidden layers with ReLu activations, and a final dense layer with output dimension of 3 and sigmoid activation, to ensure that the output corresponds to an actual color.

Type	**# filters**	**Kernel size**	**# parameters**
Conv	128	3*3	87 680
ELU	-	-	0
Batch normalization	-	-	512
Skip-co. block	-	3*3	591 744
Skip-co. block	-	3*3	591 744
Skip-co. block	-	3*3	591 744
Skip-co. block	-	3*3	591 744
Skip-co. block	-	3*3	591 744
Skip-co. block	-	3*3	591 744
Skip-co. block	-	3*3	591 744
Skip-co. block	-	3*3	591 744
Conv	5*5*3	3*3	86 475
Sigmoid activation	-	-	0
Total	-	-	4 908 619

TABLE I: **Architecture of the reconstruction network.**

B. *Reconstruction neural network*

We use a convolutional neural network with skip connections (ResNet [14]) as our reconstruction model. It consists of a stack of identical skip connection blocks, as depicted in

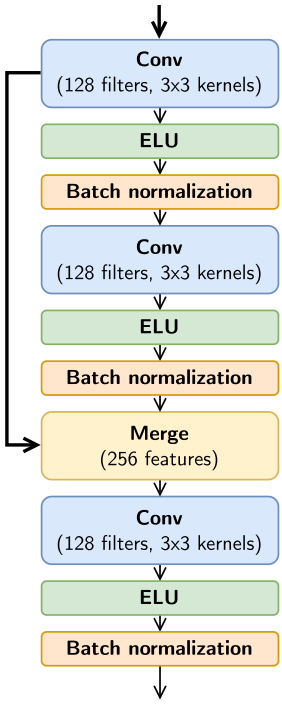


Fig. 3: **Skip connection block architecture.** It is composed of three (convolution layer + exponential linear unit activation + batch normalization) blocks. Unlike the standard residual connection that adds the input filters to some intermediate filters, the last block takes the concatenation of the input and output filters of the second block as input.

figure 3. The architecture of the reconstruction neural network is detailed in Table I.

C. End-to-end learning of the mask generation and light field reconstruction pipeline

The end-to-end learning of the complete mask generation, acquisition and reconstruction architecture depicted in figure 2 is performed by minimizing the L^1 distance between the ground truth input light fields and the light fields reconstructed by the network

$$L = \sum_{\boldsymbol{x},\boldsymbol{u},\lambda} |\mathcal{L}_{\text{g.t.}}(\boldsymbol{x},\boldsymbol{u},\lambda) - \mathcal{L}_{\text{reconstructed}}(\boldsymbol{x},\boldsymbol{u},\lambda)|$$

The end-to-end architecture is fully differentiable, which makes it possible to train everything by gradient back-propagation. No additional regularization term was used.

V. Experimental Results

We assessed the proposed framework assuming both a monochrome sensor and a sensor with a built-in CFA. We also considered both a fixed distribution for the color-coded mask and a learned one with the proposed color coded mask generator.

We compare our results with those obtained with the dictionary-based reconstruction method of [31] and with the solution proposed in [29] and using monochromatic color filter arrays. Note that the authors in [29] considered three different distributions for the color coded mask:

- uniform distribution in $[0, 1]^3$;
- either red, green or blue with equal probability (RGB);
- either red, green, blue or white with equal probability (RGBW).

since they found that the RGBW distribution performed best, we are comparing with their results using the RGBW mask. We found that the proposed reconstruction network indeed gave similar performances for both the RGBW and the uniform distributions.

A. Data sets

We evaluate the proposed framework using natural light fields from the Stanford Lytro Light Field Archive [33] and the Lytro light field data set provided by Kalantari et al. [18]. These natural light fields have been captured using the Lytro Illum camera and have an angular resolution of 14×14 and 8×8, respectively. Since these light fields suffer from strong vignetting effects on peripheral views due to mechanical and optical imperfections, we only take into account the 5×5 central angular images as reference (a.k.a ground truth) in our experiments.

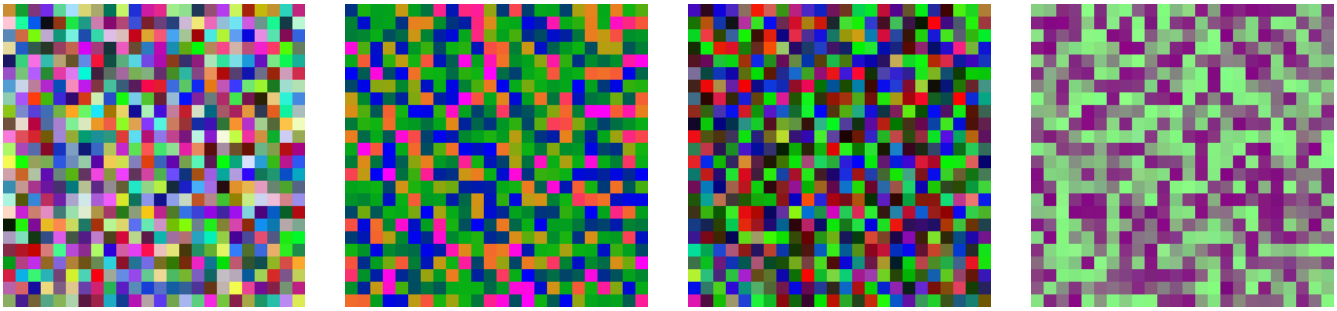
All models were trained on 122 light fields. Among these 122 light fields, 100 were originally provided by [18] as a training dataset (72 are original light fields from [18], 28 original from the Stanford dataset [35]). The remaining 22 light fields were originally provided for validation by [18] and we removed two light fields that were too similar to the *Orchids* light field, since it is conventionally used for comparison with other methods. We instead used 14 light fields from Linköping University (these light fields will be made available online soon) as our validation set.

B. Training details

While we could train all networks from scratch in a straightforward manner, we used instead the training flow illustrated in Figure 5. Fine-tuning a reconstruction network pre-trained for another task not only gives a reduced training time but also gives better performances.

All models were trained using the Adam optimizer [20] with the hyper-parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and with a numerical stability term $\epsilon = 10^{-5}$ for all four tasks using batches of size 16. For the initial task (*i.e.* monochromatic sensor + uniform CCM distribution) we used an initial learning rate $\alpha = 5 \cdot 10^{-3}$ and reduced it by a factor $10^{1/4}$ every 30 epochs. We first trained the model for 210 epochs. We also found it beneficial to train for 120 additional epochs after the network had converged, restarting from a fresh optimizer (*i.e.* one that needs to gather the statistics again) with the same initial learning rate and the learning rate decay schedule.

The remaining three tasks were all trained using an initial learning rate of $\alpha = 5 \cdot 10^{-4}$ for 60 epochs, reducing the learning rate by a factor $10^{1/4}$ every 10 epochs. Following a standard procedure, we trained on light field patches of size 100×100 in the spatial domain. Due to the full translational

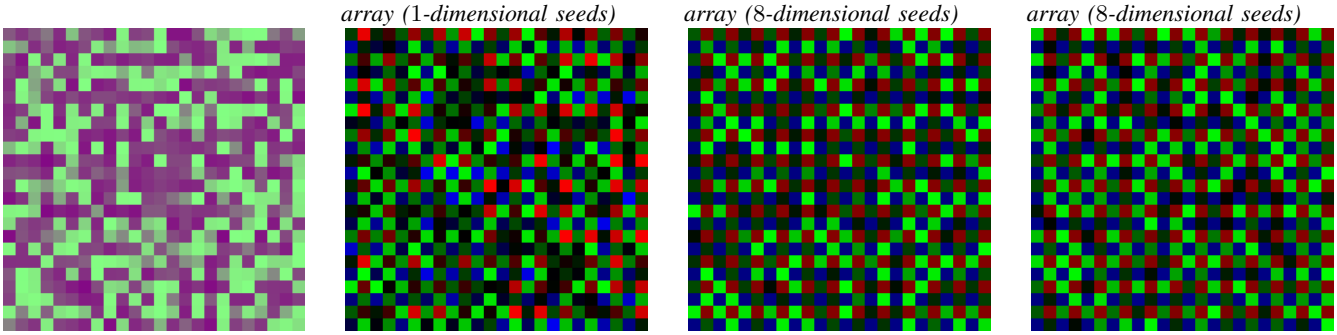


(a) *Color coded mask following uniform distribution*

(b) *Learned color coded mask distribution with no color filter array (1-dimensional seeds)*

(c) *Learned color coded mask distribution with no color filter array (8-dimensional seeds)*

(d) *Learned color coded mask distribution with color filter array (8-dimensional seeds)*



(e) *Learned color coded mask distribution with color filter array (1-dimensional seeds)*

(f) *Effective modulating tensor based on a uniform CCM*

(g) *Effective modulating tensor with the learned CCM (1-dimensional seeds)*

(h) *Effective modulating tensor with the learned CCM (8-dimensional seeds)*

Fig. 4: **Color coded masks (CCM) and effective tensors obtained by element-wise multiplication of the CCM and of the CFA.**

Method	(i)	(ii)	(ii)†	(iii)	(iv)	(iv)†	(v)†	Dict-TV[31]	Deep[29]	DeepSAS [11]
Buttercup	31.53	31.71	31.76	32.06	32.37	32.40	**32.41**	29.64	29.98	30.09
Cars	30.85	30.82	31.04	30.86	31.15	31.18	**31.22**	27.12	29.88	27.40
Orchids	32.23	32.45	32.46	32.53	32.92	**32.99**	32.89	28.41	30.99	28.67
Rock	31.06	31.08	31.17	30.83	31.15	31.20	**31.22**	27.15	30.11	30.62
Seahorse	33.45	33.45	33.52	33.21	33.33	33.33	**33.61**	29.92	32.36	29.64
Tulips	41.01	41.19	41.75	42.16	42.13	42.89	**42.89**	40.83	38.26	41.61
White rose	32.81	32.82	32.88	32.54	32.95	32.92	**32.96**	28.23	31.84	26.18
Average	33.28	33.36	33.51	33.46	33.71	33.85	**33.89**	30.19	31.92	31.78

TABLE II: **PSNR (dB) comparison with a dictionary-based reconstruction method [31] and with reference deep learning based methods [29] and [11], using a one shot acquisition scheme**. (i) Monochromatic sensor with random uniform color-coded mask distribution. (ii) Monochromatic sensor with learned color-coded mask distribution. (iii) Bayer CFA with random uniform color-coded mask distribution. (iv) Bayer CFA with learned color-coded mask distribution. (v) Learned CFA with learned color-coded mask distribution. The symbol † indicates the use of 8-dimensional seeds, while no symbol indicates a 1-dimensional seed.

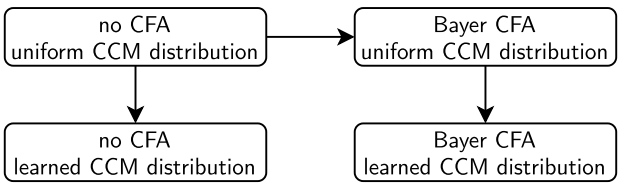


Fig. 5: **Training flow between the different tasks.** An arrow between two tasks indicates that the training of the end-to-end architecture for the target task was done using a reconstruction network pre-trained on the source task.

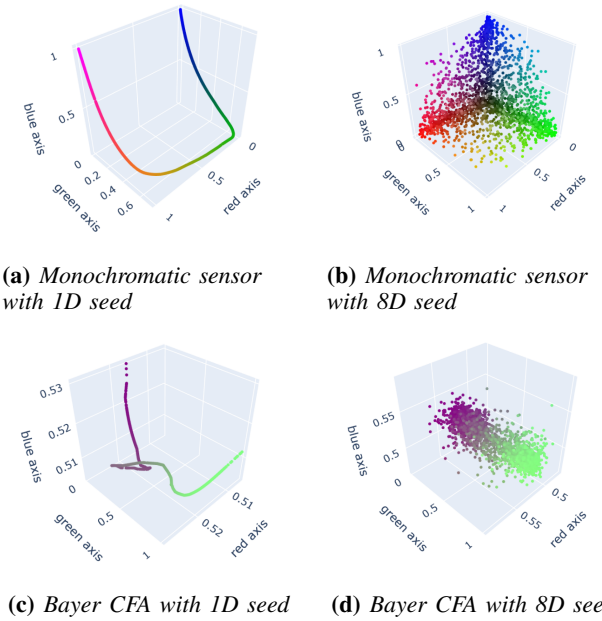
equivariance of the end-to-end architecture, it is then straightforward to use the trained network on larger light fields. We also applied on-the-fly data augmentation on each patch by randomly changing the hue in the interval $[-0.1, 0.1]$, as well as changing the saturation by a factor randomly chosen in $[0.75, 1.5]$.

C. Comparison of effective modulating tensors

Figure 4 shows a collection of color-coded masks and effective modulating tensors (calculated by element-wise multiplication of the color coded mask and of the color filter array) that we obtained and/or used in our experiments. Figure

Method	(i)	(ii)	(ii)†	(iii)	(iv)	(iv)†	(v)†	Deep[29]	DeepSAS[11]
Buttercup	0.9371	0.9362	0.9358	0.9378	0.9413	0.9413	**0.9425**	0.9216	0.9073
Cars	0.9604	0.9593	0.9596	0.9586	0.9607	0.9612	**0.9617**	0.9539	0.9230
Orchids	0.9653	0.9644	0.9646	0.9643	0.9659	0.9668	**0.9672**	0.9571	0.9388
Rock	**0.9294**	0.9283	0.9285	0.9227	0.9270	0.9275	0.9279	0.9238	0.9250
Seahorse	0.9716	0.9691	0.9707	0.9700	0.9697	0.9711	**0.9724**	0.9661	0.9496
Tulips	0.9788	0.9775	0.9782	0.9814	0.9821	**0.9832**	0.9830	0.9719	0.9802
White rose	**0.9583**	0.9579	0.9572	0.9538	0.9572	0.9568	0.9577	0.9516	0.9069
Average	0.9573	0.9561	0.9564	0.9555	0.9577	0.9583	**0.9589**	0.9494	0.9395

TABLE III: **SSIM comparison with reference deep learning based methods [29] and [11], using a one shot acquisition scheme**. (i) Monochromatic sensor with random uniform color-coded mask distribution. (ii) Monochromatic sensor with learned color-coded mask distribution. (iii) Bayer CFA with random uniform color-coded mask distribution. (iv) Bayer CFA with learned color-coded mask distribution. (v) Learned CFA with learned color-coded mask distribution. The symbol † indicates the use of 8-dimensional seeds, while no symbol indicates a 1-dimensional seed.



(a) *Monochromatic sensor with 1D seed* **(b)** *Monochromatic sensor with 8D seed*

(c) *Bayer CFA with 1D seed* **(d)** *Bayer CFA with 8D seed*

Fig. 6: **Representation of the learned distributions over $[0,1]^3$ for different configurations**. The distribution is visualized by sampling points from the output of the per-pixel generator. Points are displayed at coordinates (r,g,b) with the corresponding color. The distributions of colors for 1-dimensional generator seeds clearly lie on a 1-dimensional manifold, while the distributions for 8-dimensional generator seeds can span the whole color cube.

6 represents the corresponding learned distributions over the color space.

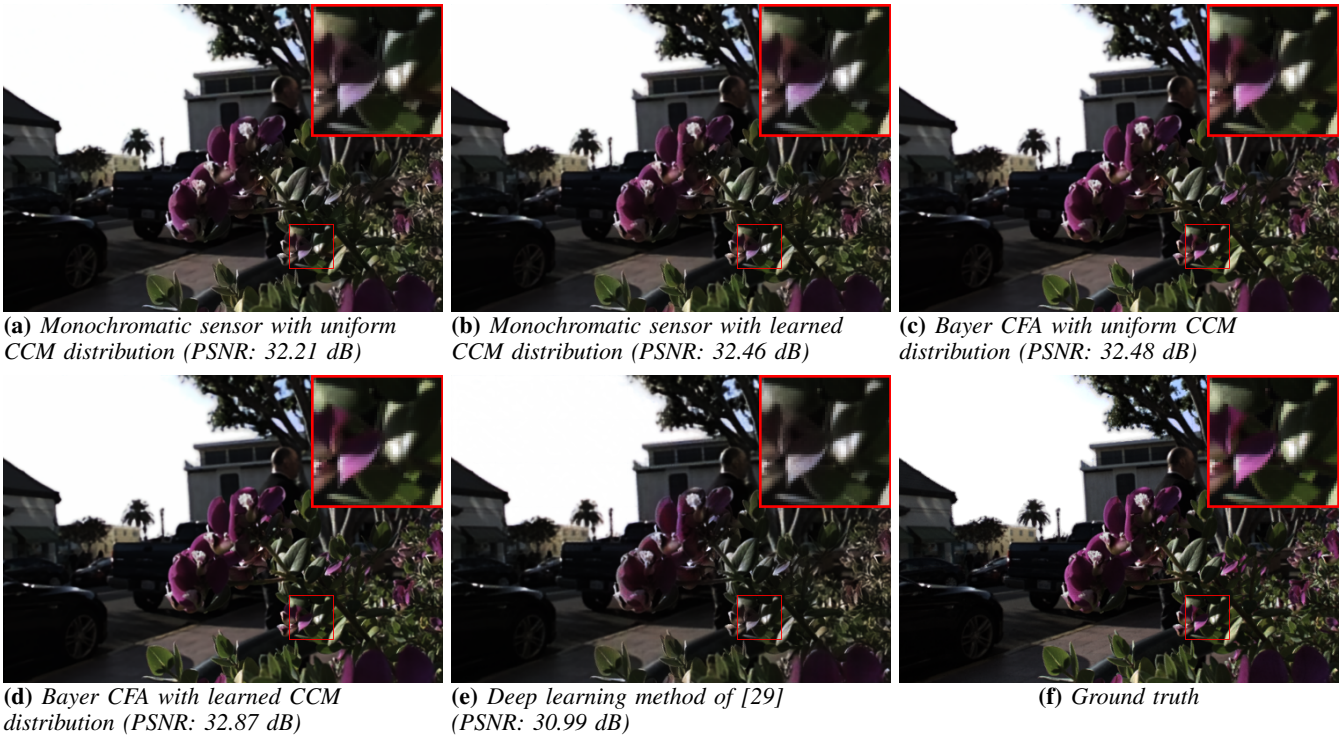
Figures 4b and 4c shows the aspect of the learned CCM in the case of a monochromatic sensor for 1D and 8D seeds, respectively. One can note that the 1D mask shows less variability in the colors than the 8D mask. This can be explained by the dimension of the support of the distribution (smaller in the 1D case), which can be visualized in Figures 6a and 6b. In both cases, however, the generated colors can span a wide range of hues.

Figures 4e and 4d show the appearance of the learned CCM in the case of a Bayer CFA for 1D and 8D seeds respectively. These masks are almost indistinguishable, even though figures 6c and 6b show that the support in the 1D case is 1-dimensional, while in the 8D case it is 3-dimensional. In fact, even though the model for the 8D color generator has enough capacity to span the entire cube, the learned distribution practically fits on a 1-dimensional manifold, showing a very distinct single degree of freedom. This is in contrast with the case with monochromatic sensor. Our hypothesis is that in the case of no CFA, the task of color compression is performed only by the CCM (in addition to the angular compression), while in the Bayer case, the CFA will help to perform the color-compression too; thus, reducing the need for the CCM to span a large portion of the color cube and present a diversity of hue.

Section IV-A discusses these phenomena in more details.

D. Comparison to the state of the art

Tables II and III present a detailed comparison of the various reconstructions obtained using our method with two reference methods, one method using a classical dictionary-based reconstruction followed by a total variation regularization (called Dict-TV) in [31] and the deep learning method of [29]. Note that although the model in [31] allows for any number of acquisitions (multiple shots), for the sake of the comparison here we consider only one shot acquisition. Moreover, note that since the authors in [29] show that the RGBW mask gives superior results with their architecture, this mask is used for the results of [29] reported here. To be able to compare to [29], all the light fields were cropped on the right and bottom after reconstruction to be of spatial size 350×500. Since our color-coded mask is non-deterministic, we take the mean of the PSNR over 32 independent realizations. The estimated standard deviation is always < 0.06 dB. We also include a comparison with the method of [11], using the implementation provided by the authors. Since the trained model provided by the authors solves the task of reconstructing a 7x7 light field from 2 shots, we had to retrain their model for solving the task $1 \rightarrow$ 5x5. We used the default hyper-parameters provided in their implementation, but used the same training dataset we used for training our model.



(a) *Monochromatic sensor with uniform CCM distribution (PSNR: 32.21 dB)*

(b) *Monochromatic sensor with learned CCM distribution (PSNR: 32.46 dB)*

(c) *Bayer CFA with uniform CCM distribution (PSNR: 32.48 dB)*

(d) *Bayer CFA with learned CCM distribution (PSNR: 32.87 dB)*

(e) *Deep learning method of [29] (PSNR: 30.99 dB)*

(f) *Ground truth*

Fig. 7: **Reconstructed central view of the *Orchids* light field using the proposed pipeline in comparison with the method of [29].** Note how the use of Bayer CFA allows for a more accurate recovery of colors, especially in high-saturation areas (see zoom-in; colors better visualized on a monitor). PSNR is indicated for the whole light field, not only the central view.

A comparison of the quantitative metrics in Table II shows that the proposed pipeline with the novel reconstruction network, the learned mask distribution, and the color filter array leads to superior performance with respect to both [31] and [29]. Qualitative results for a variety of configurations for our method using two selected light fields, namely ***Orchids*** and ***Tulips***, are shown in figures 7 and 8, respectively. We observe a better reconstruction quality with regards to the color information when compared with the results of [29]. Note that the superiority of using a Bayer CFA over a monochrome one is especially visible on colorful light fields like ***Orchids*** and ***Tulips***. Figure 9 also shows that the structures in the epipolar plane images are well preserved, hence the parallax is well reconstructed. Videos of reconstructed light fields, affirming the faithful parallax reconstruction, can be accessed at http://clim.inria.fr/research/DeepLFCam/index.html.

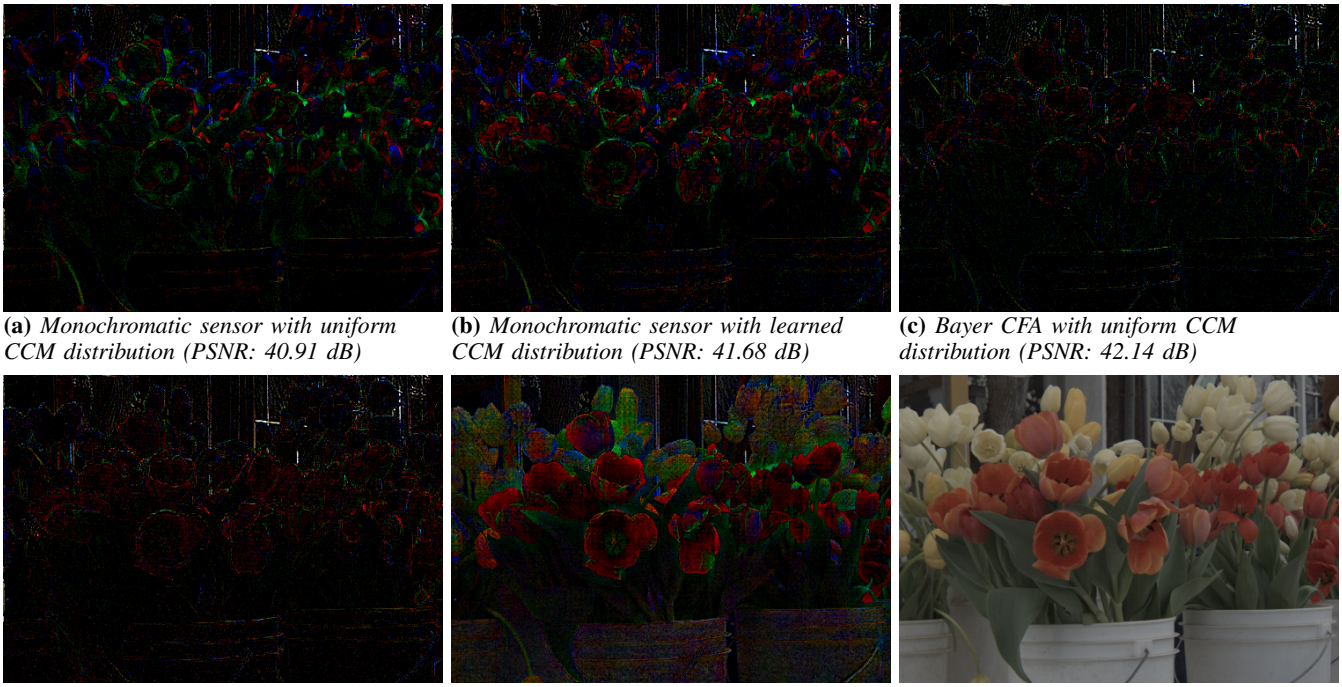
To further assess the robustness of our model, we also applied the proposed acquisition and reconstruction pipeline to synthetic light fields of the HCI dataset [39]; the results are summarized in figures 11 and 10. We would like to emphasize the fact that the light fields in the training set, which are captured using a Lytro camera, are significantly distinct to those of the HCI dataset. When comparing reconstructed central views and EPIs of synthetic light fields, as shown in Figures 10 and 11, one can see that the learned model performs well for light fields that are very different from the training ones. We invite the reader to view the animated reconstructed light fields at http://clim.inria.fr/research/DeepLFCam/index.html.

E. Ablation study

1) Proposed reconstruction network: While inspired by the reconstruction network of [29], our reconstruction networks differ in the use of skip connections in the same manner as ResNet [14]. When comparing column (i) with the last column of Table II, one can see that our architecture, even without a CFA and with a uniform coded-mask, is superior on each of the test light fields, with an average PSNR gain of 1.36dB.

2) Use of the color filter array: Table II shows the impact of the CFA and the CCM on the reconstruction quality. It can be seen that using a Bayer CFA instead of a monochromatic one contributes to improving the PSNR, in both cases of a uniform CCM or of a learned one. In addition to these quantitative results, we found that using a Bayer CFA instead of a monochromatic sensor was beneficial in terms of color reconstruction quality, while maintaining the parallax reconstruction. Table II, and figures 7 and 8, give a more detailed comparison of the different configurations of our pipeline, making apparent the benefits of using a Bayer CFA for color reconstruction.

3) Joint learning of the CFA: In addition to the learning of the coded mask, we also conducted experiments to jointly learn the color filter array. In our experiments, we learn a 4×4 pattern initialized with random values following a uniform distribution $\mathcal{U}(0, 1)$. At each training step the values are clipped to remain between 0 and 1, ensuring the physical feasibility of the CFA. Column (v)† of table II shows that jointly learning the CFA yields better results than using either a fixed monochrome sensor or a Bayer CFA, although



(a) *Monochromatic sensor with uniform CCM distribution (PSNR: 40.91 dB)*

(b) *Monochromatic sensor with learned CCM distribution (PSNR: 41.68 dB)*

(c) *Bayer CFA with uniform CCM distribution (PSNR: 42.14 dB)*

(d) *Bayer CFA with learned CCM distribution (PSNR: 43.02 dB)*

(e) *Deep learning method of [29] (PSNR: 38.26 dB)*

(f) *Ground truth*

Fig. 8: **Error maps of the reconstructed central view of the *Tulips* light field using the proposed pipeline in comparison with the method of [29].** The error maps are scaled by a factor 10 for visibility purpose. PSNR is indicated for the whole light field, not only the central view.

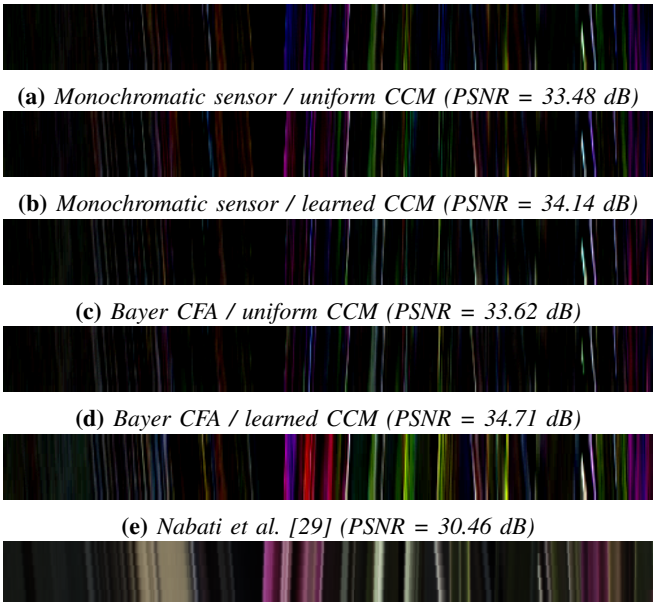
Noise	Low level				High level			
Learned CFA	No		Yes		No		Yes	
Entropy regularizer	No	Yes	No	Yes	No	Yes	No	Yes
Buttercup	31.64	30.57	**32.08**	32.04	28.54	29.26	29.54	**30.67**
Cars	30.69	30.16	30.88	**31.00**	25.73	28.50	26.06	**29.95**
Orchids	32.18	31.75	32.63	**32.73**	28.25	30.09	29.70	**31.48**
Rock	30.51	29.86	**31.07**	31.04	28.69	28.57	28.72	**30.02**
Seahorse	32.28	32.23	32.78	**33.06**	27.44	30.09	28.93	**31.63**
Tulips	40.66	40.39	40.50	**41.77**	37.33	37.57	38.77	**39.07**
White rose	32.00	31.64	32.47	**32.66**	24.95	29.55	26.74	**31.61**
Average	32.85	32.37	33.20	**33.47**	28.70	30.52	29.78	**32.06**

TABLE IV: **PSNR (dB) comparison of different models under noisy measurements.** The first four columns correspond to $g = 1$ (low noise level), while the last four correspond to $g = 0.33$ (high noise level). Configurations not learning the CFA use a fixed Bayer CFA instead. Boldface figures indicate the best performing model for the corresponding level of noise.

improving over the Bayer CFA by less than 0.1 dB. However the effectiveness of learning the CFA becomes much more visible in presence of noise as shown in Section V-F, yielding improvement of more than 1 dB compared to using a fixed Bayer CFA. Figure 12 shows color filter arrays corresponding to various levels of noise.

4) d-dimensional seed versus 1-dimensional seeds: Since the color distribution of the pixels of the CCM is learned by training a continuous mapping from $\mathbf{R}^d$ to $[0,1]^3$, the support of the distribution is a manifold whose dimension is at most d. If for instance we choose $d = 1$, the support of the distribution will be a continuous and piece-wise differentiable curve within the color cube. No such mapping can therefore perfectly fit distributions whose support is 2D or 3D. It is, nonetheless, possible to approximate any distribution, but at the expense of an arbitrarily complex winding curve. For this reason, we expect generators with high-dimensional seeds to perform better in general. A visual representation of this fact is presented by Figure 6. We empirically found that using an 8D seed, instead of a 1D one, leads to an increase in PSNR of about 0.15dB.

5) Impact of CCM resolution: The initial design assumes that the coded mask $M(\boldsymbol{\xi}, \lambda)$ has a feature size $\triangle_M$, *i.e.* the size of each element or pixel of the mask, and that it is placed at a distance d from the sensor plane ($0 \leq d \leq D$) in a way that $\frac{d}{\triangle_M}$ is sufficiently small with respect to the visible spectrum



(a) *Monochromatic sensor / uniform CCM (PSNR = 33.48 dB)*

(b) *Monochromatic sensor / learned CCM (PSNR = 34.14 dB)*

(c) *Bayer CFA / uniform CCM (PSNR = 33.62 dB)*

(d) *Bayer CFA / learned CCM (PSNR = 34.71 dB)*

(e) *Nabati et al. [29] (PSNR = 30.46 dB)*

(f) *Ground truth*

Fig. 9: **Visualization of the error of epipolar plane images (EPI) of the *Orchid* light field**. These are obtained by spanning the horizontal middle line of the light field. That is, displaying the mapping $(\omega_x, \nu_u, \lambda) \mapsto \mathcal{L}(\omega_x, \omega_y^{\text{middle}}, \nu_u, \nu_v^{\text{middle}}, \lambda)$. Error maps have been amplified by a factor 6 for better visualization. PSNR indicated with respect to the EPI.

CCM pixel size	1	2	4	8
PSNR	33.47	**33.63**	33.62	33.38
SSIM	0.9555	**0.9568**	0.9568	0.9555

TABLE V: **Performance comparison between different resolutions of the coded mask.** The size of the CCM's pixels is indicated relative to the size of a the sensor's pixels. For all configurations the sensor is equipped with a Bayer CFA, and the CCM's pixels follow a uniform distribution.

(corresponding to the wavelength in the range between 0.4 and 0.8 μm). Under these conditions, the diffraction effect induced by the mask M can be ignored [2]. However, we have further assessed the impact of the CCM resolution, *i.e.* of the size of the CCM elements, on the reconstructed light fields. In that case, the reconstruction problem becomes a joint problem of reconstruction and de-convolution. Table V shows the impact of the CCM resolution on the reconstruction. One remarks that diminishing the resolution of the CCM can, up to a certain point, improve the performance of the reconstruction network. One explanation to this seemingly counter-intuitive fact is that, as we reduce the resolution, we also reduce the variety of patterns to which the reconstruction network can better adapt. We also hypothesize that the presence of the Bayer CFA makes the system more robust to a lower resolution of the CCM, as some color multiplexing is still performed at the pixel level. Decreasing too much the CCM resolution hurts the performance, as it makes the sensing matrix less incoherent.

F. Noise Analysis

1) Noise model: In order to assess the robustness of our method under practical conditions, we have corrupted the coded projections with sensor random noise. Our corruption model takes into account *shot noise* due the quantized nature of light, *readout* noise caused by spontaneous emission of electrons in the sensor, and rounding happening during analog-digital conversion. Denoting $\mathcal{I}$ the normalized ground truth intensity (*i.e.* proportional to the irradiance) of a pixel, the normalized intensity measured by the sensor is given by:

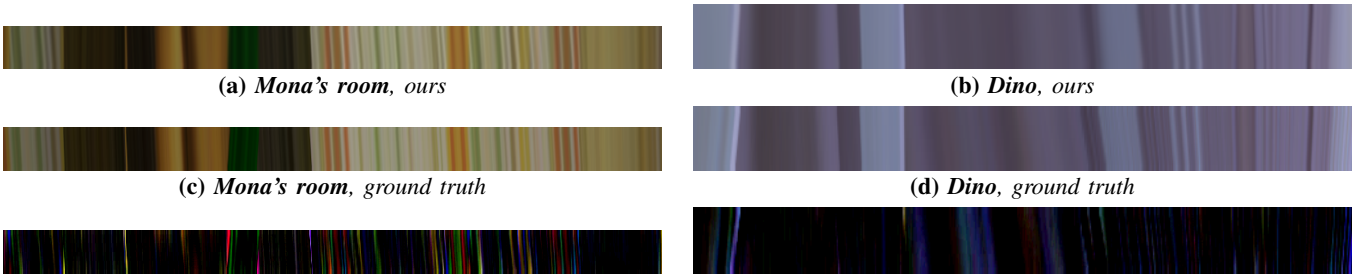
$$\mathcal{I}_{\text{sensor}} = \frac{1}{gN}\left[\frac{N}{c}\text{clip}_{[0,c]}(\mathbf{p}(gc\mathcal{I}) + \mathbf{n}_{\text{readout}})\right]$$

where $[.]$ denotes rounding, $\text{clip}_{[0,c]}(.) = \max(\min(., c), 0)$, $N = 2^b - 1$, with b the number of bits used to code the digital converted measure, c is the full well capacity of the pixel (in number of electrons), g is a gain factor proportional to the ISO gain of the pixel and the exposure time, $\mathbf{p}(gc\mathcal{I})$ is a random variable following a Poisson distribution $\mathcal{P}(gc\mathcal{I})$ and $\mathbf{n}_{\text{readout}} \sim \mathcal{N}(0, \sigma_{\text{readout}})$.

However, with this model, $\mathcal{I}_{\text{sensor}}$ is not differentiable with respect to $\mathcal{I}$, because of the rounding operator and the sampling of the Poisson distribution. This non-differentiability prevents the back-propagation of gradients to elements upstream of the sensing operator. In order to make it differentiable, we replaced the rounding operation by a uniform additive noise $\mathbf{n}_{\text{rounding}} \sim \mathcal{U}(-1/2, 1/2)$ and $\mathbf{p}(gc\mathcal{I})$ by a Gaussian variable $gc\mathcal{I} + \sqrt{gc\mathcal{I}} \cdot \mathbf{n}_{\text{shot}}$ where $\mathbf{n}_{\text{shot}} \sim \mathcal{N}(0, 1)$. Such substitutions preserve the order of magnitude of the corruption applied to the signal, while allowing to back-propagate gradients up to the CFA and CCM.

In our experiments, we set $\sigma_{\text{readout}} = 40$, $N = 2^{14} - 1$ and $c = 20000$, which are typical parameters corresponding to a medium-low quality photosensor. We consider two values of gain: $g = 1$ and $g = 0.33$, which correspond on our dataset to noise levels of about 1.5% and 3% respectively. We experimented with both a fixed Bayer CFA and a trainable CFA. Table IV shows that our model is robust to noise, when the noise level is not too high. It also demonstrates that learning the CFA further increases the reconstruction quality in presence of noise. When the noise level is high, to avoid the harmful impact of the noise on the learned CCM and preserve a good robustness of the approach, we introduce, as explained below, an entropy-based regularization of the CCM.

2) Entropy regularization of the coded mask: We noticed that, as we increase the level of noise, the color distribution learned by the coded mask tends to become less diverse. With moderate noise levels, the coded mask rapidly converges to a binary mask, both with a fixed Bayer CFA and a learned CFA. When increasing the noise level, the coded mask converges to a uniform transparent mask. One possible explanation is that because of the noise, the coded mask learns very early to increase its light efficiency and falls into a local minimum where the CCM distribution is very simple, because the task of reconstructing from simply multiplexed measurements is easier than the task of reconstructing from incoherent measurements, since the reconstruction network has to learn how to make use of the sensing matrix. In order to overcome this



(a) ***Mona's room****, ours*

(b) ***Dino****, ours*

(c) ***Mona's room****, ground truth*

(d) ***Dino****, ground truth*

(e) ***Mona's room****, error map (EPI PSNR = 35.47 dB)*

(f) ***Dino****, error map (EPI PSNR = 35.72 dB)*

Fig. 10: **Visualization of epipolar plane images (EPI) of the *Mona's Room* and *Dino* synthetic light fields from the HCI dataset [39]**. Reconstructed using the Bayer CFA and learned CCM distribution architecture. The error maps as scaled by a factor 10.

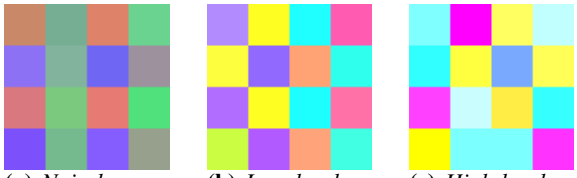
(a) *"Mona's room", ours (PSNR: 34.72 dB)*

(b) *"Mona's room", ground truth*

(c) *"Dino, ours" (PSNR: 33.78 dB)*

(d) *"Dino", ground truth*

Fig. 11: **Reconstructed central view of the *Mona's Room* and *Dino* light fields from the HCI dataset [39]** using the proposed pipeline with Bayer CFA and learned CCM architecture.



(a) *Noiseless acquisition (light efficiency = 60%)*

(b) *Low-level noise (light efficiency = 70%)*

(c) *High-level noise (light efficiency = 78%)*

Fig. 12: **Learned color filter array patterns for different levels of noise.** Note how the light efficiency increases with the level of noise.

problem, we encouraged variety in the CCM distribution by adding a novel regularization term on the output of the CCM generator network. This regularization term is given by:

$$R_{\text{CCM}} = \max(0, \mathcal{H}_{\text{threshold}} - \hat{\mathcal{H}}_{\text{KL}}(c_1, \ldots, c_n))$$

where $c_1, \ldots, c_n$ are n color values sampled from the CCM generator, $\hat{\mathcal{H}}_{\text{KL}}(.)$ is the Kozachenko-Leonenko entropy estimate [21], and $\mathcal{H}_{\text{threshold}}$ is a threshold value. This regularizer effectively enforces the coded mask distribution to maintain an entropy above the threshold $\mathcal{H}_{\text{threshold}}$. The Kozachenko-Leonenko estimator is a continuous, piece-wise differentiable estimator [24], and is easily implementable in modern deep learning frameworks, making it suitable for gradient-based learning with automatic differentiation. Table IV shows the effectiveness of the regularization, especially in the case of strong noise where it yields an average improvement of about 2 dB.

VI. Conclusion

We have presented an end-to-end learning framework for compressive light field acquisition with coded masks. While the state-of-the-art method only considers monochrome sensors, we also consider sensors with in-built CFA and we have show its impact on the reconstruction quality. Compared to state-of-the-art methods, our framework also learns the distribution of the colour coded mask, which is shown to outperform masks based on fixed uniform distributions. The reconstruction network using skip connections is also shown to outperform the reference architecture in the same acquisition conditions.

In addition to the learning of the coded mask distribution, we have considered the joint learning of the CFA, whose importance has been demonstrated in this paper by the superior quality of color reconstruction of the pipeline. A possible avenue for research could also be the extension of our work to multi-shot light field acquisition, which has been shown to greatly improve the reconstruction quality [31]. Since we learn a distribution for the mask, as opposed to a fixed learned mask, multi-shot acquisition can be achieved by merely modifying the reconstruction network that takes multiple realizations of the mask distribution as input.

References

[1] E. Arias-Castro, E. J. Candes, and M. A. Davenport. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2013.

[2] M. S. Asif, A. Ayremlou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk. FlatCam: Thin, Lensless Cameras Using Coded Aperture and Computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 2017.

[3] S. D. Babacan, R. Ansorge, M. Luessi, P. Ruiz Mataran, R. Molina, and A. K. Katsaggelos. Compressive Light Field Sensing. *IEEE Transactions on Image Processing*, 21(12):4746–4757, Dec 2012.

[4] S. D. Babacan, R. Ansorge, M. Luessi, R. Molina, and A. K. Katsaggelos. Compressive sensing of light fields. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 2337–2340, Nov 2009.

[5] B. E. Bayer. Color Imaging Array, Jul 1976.

[6] A. Bora, A. Jalal, E. Price, and A. G. Dimakis. Compressed sensing using generative models. In *ICML*, 2017.

[7] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2005.

[8] A. Chakrabarti. Learning sensor multiplexing design through back-propagation. In *International Conference on Neural Information Processing (NIPS)*, page 3089–3097, Dec 2016.

[9] L. H. Chang and J. Y. Wu. An improved rip-based performance guarantee for sparse signal recovery via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 60(9):5702–5715, September 2014.

[10] S. Elmalem, R. Giryes, and E. Marom. Learned phase coded aperture for the benefit of depth of field extension. *Optic Express*, 26(12):15316–15331, 2018.

[11] M. Guo, J. Hou, J. Jin, J. Chen, and L.-P. Chau. Deep spatial-angular regularization for compressive light field reconstruction over coded apertures. In *European Conference on Computer Vision (ECCV)*, pages 278–294, Oct 2020.

[12] M. Gupta, A. Jauhari, K. Kulkarni, S. Jayasuriya, A. Molnar, and P. Turaga. Compressive light field reconstructions using deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1277–1286, Jul 2017.

[13] H. Haim, S. Elmalem, R. Giryes, A. M. Bronstein, and E. Marom. Depth estimation from a single image using deep learned phase coded mask. *IEEE Transactions on computational imaging*, 4(3):298–310, 2018.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[15] B. Henz, E. Gastal, and M. Oliveira. Deep joint design of color filter arrays and demosaicing. *Computer Graphics Forum*, 37, 05 2018.

[16] M. Hirsch, S. Sivaramakrishnan, S. Jayasuriya, A. Wang, A. Molnar, R. Raskar, and G. Wetzstein. A switch-

able light field camera architecture with angle sensitive pixels and dictionary-based sparse coding. In *International Conference on Computational Photography (ICCP)*, pages 1–10, 2014.

[17] Y. Inagaki, Y. Kobayashi, K. Takahashi, T. Fujii, and H. Nagahara. Learning to capture light fields through a coded aperture camera. In *The European Conference on Computer Vision (ECCV)*, Sep 2018.

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

[19] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pat- tern Recognition (CVPR)*, pages 1646–1654, June 2016.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[21] L.F. Kozachenko and N.N. Leonenko. Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, 23(2):95–101, 1987.

[22] M. Levoy and P. Hanrahan. Light Field Rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH, pages 31–42. ACM, 1996.

[23] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen. Programmable Aperture Photography: Multiplexed Light Field Acquisition. *ACM Transactions on Graphics*, 27(3):55:1–55:10, Aug 2008.

[24] D. Lombardi and S. Pant. Nonparametric k-nearest-neighbor entropy estimator. *Physical Review E*, 93(1):013310, 2016.

[25] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar. Compressive Light Field Photography using Overcomplete Dictionaries and Optimized Projections. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, 32(4):46:1–46:12, 2013.

[26] E. Miandji, J. Kronander, and J. Unger. Compressive Image Reconstruction in Reduced Union of Subspaces. *Computer Graphics Forum*, 34(2):33–44, May 2015.

[27] E. Miandji, J. Unger, and C. Guillemot. Multi-shot single sensor light field camera using a color coded mask. In *European Signal Processing Conference (EUSIPCO)*, pages 226–230, Jun 2018.

[28] Ehsan Miandji, Mohammad Emadi, Jonas Unger, and Ehsan Afshari. On probability of support recovery for orthogonal matching pursuit using mutual coherence. *IEEE Signal Processing Letters*, 24(11):1646–1650, November 2017.

[29] O. Nabati, D. Mendlovic, and R. Giryes. Fast and accurate reconstruction of compressed color light field. *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–11, 2018.

[30] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan. Light Field Photography with a Handheld Plenoptic Camera. Computer Science Technical Report CSTR 2(11), Stanford University, 2005.

[31] H.-N. Nguyen, E. Miandji, and C. Guillemot. Multi-mask camera model for compressed acquisition of light fields. *IEEE Trans. on Computational Imaging, accepted*, 2020.

[32] Y.C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 40–44. IEEE, November 1993.

[33] A. S. Raj, M. Lowney, R. Shah, and G. Wetzstein. Stanford Lytro Light Field Archive. http://lightfields.stanford.edu/LF2016.html, Oct 2016. Online; released Oct 2016.

[34] V. Sitzmann, S. Diamond, Y. Peng, X. Dun, S. Boyd, W. Heidrich, F. Heide, and G. Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics*, 37(4):114:1–114:13, July 2018.

[35] Stanford. Stanford lytro light field archive, 2018.

[36] J.A Tropp. Greed is good: algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10):2231–2242, October 2004.

[37] A. K. Vadathya, S. Cholleti, G. Ramajayam, V. Kanchana, and K. Mitra. Learning light field reconstruction from a single coded image. In *Asian Conference on Pattern Recognition (ACPR)*, 2017.

[38] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin. Dappled photography: Mask enhanced cameras forheterodyned light fields and coded aperture refocusing. In *ACM SIGGRAPH 2007*, volume 26, 2007.

[39] S. Wanner, S. Meister, and B. Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *Conference on Vision, Modeling, and Visualization (VMV)*, pages 225–226, 2013.

[40] G. Wetzstein, I. Ihrke, D. Lanman, and W. Heidrich. State of the Art in Computational Plenoptic Imaging, booktitle = IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010.

[41] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, a. Barth, A. Adams, M. Horowitz, and M. Levoy. High Performance Imaging using Large Camera Arrays. *ACM Transactions on Graphics*, 24(3):765–776, Jul 2005.

[42] Z. Xu, J. Ke, and E. Y. Lam. High-resolution Lightfield Photography Using Two Masks. *Optics Express*, 20(10):10971–10983, May 2012.

[43] Z. Xu and E. Y. Lam. A high-resolution lightfield camera with dual-mask design. In *Proceedings of SPIE, Image Reconstruction from Incomplete Data VII, SPIE Optical Engineering + Applications*, volume 8500, Oct 2012.

[44] Y.Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan. Phasecam3d — learning phase masks for passive single view depth estimation. In *IEEE International Conference on Computational Photography (ICCP)*, pages 1–12, May 2019.