Deep Unrolling for Light Field Compressed Acquisition using Coded Masks

Guillaume Le Guludec, Christine Guillemot

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Deep Unrolling for Light Field Compressed Acquisition using Coded Masks

GUILLAUME LE GULUDEC [1], CHRISTINE GUILLEMOT [1], (Fellow, IEEE)

[1]Inria Rennes – Bretagne-Atlantique 263 Avenue Général Leclerc, 35042 Rennes Cedex, France (e-mail: firstname.lastname@inria.fr).

Corresponding author: Guillaume Le Guludec (e-mail: guillaume.le-guludec@inria.fr).

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.

ABSTRACT Compressed sensing using color-coded masks has been recently considered for capturing light fields using a small number of measurements. Such an acquisition scheme is very practical, since any consumer-level camera can be turned into a light field acquisition camera by simply adding a coded mask in front of the sensor. We present an efficient and mathematically grounded deep learning model to reconstruct a light field from a set of measurements obtained using a color-coded mask and a color filter array (CFA). Following the promising trend of *unrolling optimization algorithms* with learned priors, we formulate our task of light field reconstruction as an inverse problem and derive a principled deep network architecture from this formulation. We also introduce a closed-form extraction of information from the acquisition, while similar methods found in the recent literature systematically use an approximation. Compared to similar deep learning methods, we show that our approach allows for a better reconstruction quality. We further show that our approach is robust to noise using realistic simulations of the sensing acquisition process.

INDEX TERMS Light field imaging, compressed sensing, deep learning, inverse problems, algorithm unrolling

I. INTRODUCTION

LIGHT field imaging is becoming an increasingly popular subject of interest and research. Indeed, light fields naturally extend the traditional notion of images as they capture more information about a scene, by recording not only the radiance on pixels on a two-dimensional plane, but also discriminating along the direction of the light rays. Light field acquisition thus makes possible a great deal of applications, like post-capture image refocusing and depth-of-field tuning [1], [2], synthetic aperture imaging [3], microscopy [4], virtual reality, depth estimation [5], [6], etc. However, light field acquisition remains the main problem, as the devices that record light fields are generally very bulky as well as expensive, thus hindering light field capture in real-world applications. Plenoptic cameras based on micro-lenses placed between the main lens and the sensor have been designed [7]. Since the sensor resolution is limited, with plenoptic cameras a dense angular sampling is obtained at the expense of a reduced spatial sampling of the different views. This trade-off between spatial and angular resolutions needs to be addressed by using light field super-resolution methods as proposed in [6]. Among the other possible solutions to tackle the problem of practical light field acquisition, compressed light field acquisition using color-coded masks (CCM) and color filter array (CFA) is especially interesting. Indeed, compressed sensing is a mathematical framework providing strong guarantees on signal recovery from incomplete measurements [8]–[10]. Its application to the compressed acquisition of light fields can be made effective in several ways. Compressed sensing is originally meant to reconstruct a signal that is assumed to be sparse in a given dictionary. While traditional methods using sparse priors have been to some extent successfully applied to light field reconstruction [11]–[14], the iterative nature of the reconstruction algorithm greatly precludes its actual use, notably in real-time applications. Alternatively, deep learning methods have been very successful at reconstructing light fields in a way that is both fast and accurate [15], [16]. These methods are orders of magnitude faster than traditional iterative methods like the orthogonal matching pursuit [17], and can outperform sparsity priors for a great number of

image reconstruction tasks [18]. However, these networks in general have a very large number of parameters, hence require a very large amount of training data.

Unrolled optimization algorithms have emerged as efficient solutions to combine the flexibility of deep learning methods with the mathematical principles underpinning the more traditional optimization methods. Algorithm unrolling can be seen as a *paradigm* to design deep neural architectures from optimization algorithms while incorporating useful priors into the models in a principled way. Several optimization algorithms have been unrolled in the literature, such as ISTA and the coordinate descent algorithm [19], the gradient descent [20], the proximal gradient [21], the ADMM [22], [23] and the half quadratic splitting (HQS) [24] algorithms. A learned network is used at each iteration of the optimization algorithm, as a regularizer (or as its gradient [20]), or as a proximal operator [23] which can be seen as a denoiser. While usual optimization methods iterate until convergence using an ending condition on the reconstruction error, unrolling an iterative algorithm considers a small number of iterations. This allows training a learned component end-to-end within the optimization algorithm, in a way that takes into account the data term, *i.e.* , the degradation operator. By *end-to-end learning* we refer to the learning of the weights of the regularizers by back-propagation of the gradients from the last to the first iteration of the unrolled optimization algorithm.

In this paper, we present an unrolled optimization solution with a neural network-based prior for light field reconstruction from a set of 2D measurements. The proposed solution is based on unrolling the half-quadratic splitting (HQS) method, where the proximal operator used in the regularization step is defined by a neural network-based denoiser learned at each iteration of the unrolled optimization. While most compressed light field acquisition methods assume RGB measurements, our acquisition scheme does not make such assumption and relies instead on measurements obtained using a monochromatic sensor equipped with a color filter array. We show that our approach can be applied to any number of measurements and give experimental results for a number of measurements going from 1 to 3. Whereas the methods in [15], [16] are based on convolutional neural networks (CNN) that act as regression networks, we derive instead our architecture from the unrolling of the minimization of a regularized mean-squared problem. The proposed solution is actually based on an unrolled Half Quadratic Splitting (HQS) optimization method with a learned proximal operator, leading to a completely different architecture. We additionally show that, with our acquisition scheme, the extraction of information from the set of measurements is made possible in closed-form. This comes from the specific structure of our sensing matrix which is such that this matrix can be inverted in a very efficient manner using the Sherman-Morrison-Woodbury identity. The unrolled optimization method, together with the closed form introduced to solve the data term minimization, makes the architecture more mathematically grounded, compared with a learned CNN acting as a regression network between the input measurements and the reconstructed data. To the best of our knowledge, a closed-form minimization of the data term has not been proposed before for light field compressed acquisition, as other unrolled methods usually rely on an approximation instead (often a single gradient descent step). Note that, while unrolled optimization has been considered in [25], this was in a context of coded aperture acquisition of each colour component. Here, we consider an acquisition scheme with coded masks that is suitable for monochromatic sensors equipped with CFA.

We have assessed the proposed scheme, considering different acquisition scenarios, *i.e.* using a coded mask placed on the sensor in comparison with coded aperture designs. For the sake of comparison with the methods in [14], [15] and [16] which also consider masks placed on the sensor, we first consider a mask drawn from a uniform distribution, with a model assuming the presence of pinholes on the aperture plane. This corresponds to the case where the real continuous light field is first discretized and then modulated with the coded mask and the color filter array. We then derive a more realistic model without pinholes and show that this new model does not lead to any drop in quality. We also investigate the benefit of learning the distribution of a realization of the mask.

In summary, our contributions are as follows:

- We present a novel architecture for compressed light field acquisition using a color-coded mask and a CFA, that can take different numbers of shots at its input. This architecture is based on an unrolled HQS optimization method with a learned proximal operator.
- In the 1-shot case, we show that our architecture yields a 2.5 dB improvement over the deep learning based method in [16], which is already in average 1.97 dB better than the method in [15], using the same acquisition scheme. We also show that our approach yields a significant improvement in PSNR compared to traditional iterative methods using the same acquisition scheme.
- We present a closed-form data term minimization solution which not only makes the architecture more mathematically grounded, but also consistently gives a PSNR gain of about 0.6 dB, regardless of the number of shots.
- We derive a new realistic sensing model with coded masks placed on the sensor, which, unlike prior work, does not assume the presence of pinholes on the aperture plane to discretize the real continuous light field in the angular dimension before being modulated with the coded mask and the color filter array.
- We show that, in presence of sensor noise, the joint learning of the CCM mask distribution and the reconstruction network yields extra 0.81 dB and 2.12 dB gains in average, for low and high level of noise respectively.

II. RELATED WORK

Many camera designs have been proposed for light field acquisition. The goal of this section is not to give a complete overview of the various designs, which can be found in [26], but rather to recall the designs related to the proposed approach, *i.e.* . based on coded masks. While programmable aperture approaches with non refractive masks placed at the aperture have been proposed in [27] to sequentially capture subsets of light rays [27], we focus here on solutions based on coded masks placed in front of the sensor.

A. COMPRESSIVE LIGHT FIELD ACQUISITION WITH CODED MASKS

The problem of light field reconstruction from the recorded set of measurements can be placed in a compressed sensing framework. Thanks to the use of a coded mask, the photosensor records a set of linear measurements from which a higher resolution light field can be reconstructed. This problem, being an ill-posed inverse problem, is solved using a least squares minimization with some regularization constraint based on hand-crafted signal priors. Marwah *et al.* [12] propose a camera architecture that records optically coded projections on a single image sensor using a monochrome mask, while Miandji *et al.* in [28] and [13] use respectively a random stationary or a moving color-coded mask to extract incoherent measurements. Nguyen *et al.* [14] introduce an Equivalent Multi-Mask Camera (EMMC) model which allows for a flexible configuration of a variety of sensing schemes.

Whereas the camera designs in [28], [13], [14] place the mask close to the sensor, the coded aperture cameras have the mask placed directly on the aperture plane. This is the approach followed in [29], [30], [25] and [31], where incoherent light field measurements are captured by using a randomly coded mask placed on the aperture plane. In our proposed design, the mask is placed close to the sensor, which actually allows the rays coming from different angles to be multiplexed in a way that is dependent on the spatial position of the incident pixel, thereby increasing the possibility for the various measurements to be mutually incoherent, which is known to be a crucial property of degradation matrices as explained by [32].

B. DEEP COMPRESSIVE LIGHT FIELD ACQUISITION

In the above cases, the light field is reconstructed using a compressive sensing framework and classical sparse recovery methods, assuming the data to be sparse in a domain defined by an overcomplete dictionary [12], [13]. However, this problem can also be efficiently solved using deep learning techniques [15], [30], [33], [34]. The authors in [15], [33], [34] assume a pre-defined mask pattern and propose convolutional neural network architectures to reconstruct the light field from the set of measurements. Inagaki *et al.* [30] formulate the coded aperture acquisition and light field reconstruction as an auto-encoder, whereas we consider instead an unrolled optimization technique, and optimize the mask pattern together with the parameters of the reconstruction algorithm in an end-to-end auto-encoder learning. A learned convolutional network architecture is used in [25] 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.

C. UNROLLED OPTIMIZATION METHODS

Recent methods have been introduced with the goal of combining the advantages of well understood iterative optimization techniques with those of learned regularizers allowing us to model more complex image priors. These regularizers can take the form of operators of projection on a learned image subspace or manifold [18], [35], of a denoiser [36], [35], [37], or of the proximal operator for a regularizer [38].

Unrolling a fixed number of iterations of optimization algorithms is an efficient way of coupling optimization and deep learning techniques. Whereas usual iterative methods iterate until convergence using an ending condition on the reconstruction error, unrolling an iterative algorithm considers a small number of iterations. This allows using a learned regularization step trained end-to-end within the optimization algorithm, hence in a way that takes into account the data term, *i.e.* , the degradation operator. This principle has been applied to several optimization algorithms in the literature, *e.g.* , to ISTA and the coordinate descent algorithm [19], the gradient descent [20], the proximal gradient [21], the ADMM [22], [23] and HQS [24] algorithms. The learned regularization network can take the form of the gradient of a regularizer [20], or of a proximal operator based on a denoiser [23]. The number of iterations of unrolled optimization methods is typically quite small due to difficulties in training networks corresponding to a large number of iterations. The authors in [39] however propose a solution for training arbitrarily deep unrolled optimization networks based on deep equilibrium models [40].

III. MATHEMATICAL FORMULATION OF THE LIGHT FIELD ACQUISITION PROBLEM

A. LIGHT FIELDS

A light field is in general defined as the radiance at a given point in space and time, along a given direction and for a given wavelength. As such, it may be described as a real-valued function over a 7D space $L(x, y, z, \theta, \phi, \lambda, t)$, where (x, y, z) are the spatial coordinates, and (θ, ϕ) are the angular coordinates, λ is the wavelength and t is the time. In the context of static light field acquisition by a given camera, assuming free space around the camera, a light field is more adequately described by a well-known *two-plane parameterization* $L(x, y, u, v, \lambda)$ where (x, y) are the spatial coordinates, *i.e.* the coordinates on the *sensor plane*, and (u, v) are the angular coordinates, *i.e.* the coordinates on the *aperture plane*. For the sake of notation simplicity, we write $\boldsymbol{x} = (x, y)$ and $\boldsymbol{u} = (u, v)$. We define $\boldsymbol{X}$ as the range of spatial coordinates $\boldsymbol{x}$, $\boldsymbol{U}$ as the range of angular coordinates $\boldsymbol{u}$ and Λ as the range of the spectral coordinate λ.

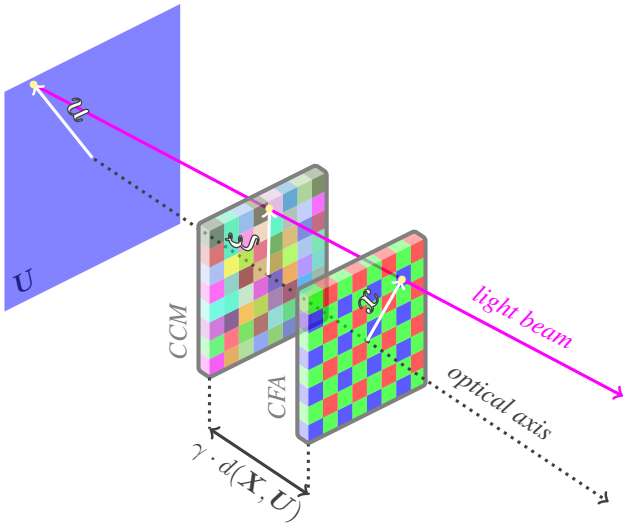


FIGURE 1: **Model of compressed light field acquisition.** A light beam, parameterized by $(\boldsymbol{x}, \boldsymbol{u})$, is first filtered by the color-coded mask at coordinates $\boldsymbol{\xi}$ on the CCM plane. It is subsequently filtered by the color filter array on the sensor plane $\boldsymbol{X}$.

B. MONOCHROME IMAGE FORMATION MODEL

We consider a general camera model in which a light field L is first linearly modulated by one or several optical devices, before being recorded by an ideal 2D photosensor. In practice, the light field is usually first transformed by a convergent lens, but this process can be ignored, without loss of generality, by considering L to be the in-camera conjugate light field with respect to the lens. For the sake of simplicity we also ignore the vignetting effect in our model, since it can be included in L as well. Assuming $\boldsymbol{X}$ is partitioned into P sub-regions $(\boldsymbol{X}_p)_{1\leq p\leq P}$, each corresponding to an actual pixel, the intensity recorded by the pth pixel is modeled as

$$I_p = \int_{\boldsymbol{X}_p\times\boldsymbol{U}\times\Lambda} L(\boldsymbol{x},\boldsymbol{u},\lambda)\psi(\boldsymbol{x},\boldsymbol{u},\lambda)\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{u}\mathrm{d}\lambda \qquad (1)$$

where ψ is a *modulation field* (or *shield field* [41]) conditioned by the optical components of the camera. The whole measurement consists in the combined intensities recorded by all pixels. We may further extend the model by allowing multiple-shot captures. In this case, we consider N modulation fields $(\psi_n)_{1\leq n\leq N}$, each corresponding to a shot. The number of measurements then becomes $N \times P$, and the intensity recorded by the pth pixel at the nth shot is:

$$I_{n,p} = \int_{\boldsymbol{X}_p\times\boldsymbol{U}\times\Lambda} L(\boldsymbol{x},\boldsymbol{u},\lambda)\psi_n(\boldsymbol{x},\boldsymbol{u},\lambda)\mathrm{d}x\mathrm{d}\boldsymbol{u}\mathrm{d}\lambda \qquad (2)$$

The whole acquisition process can be formulated as a linear operator:

$$\text{Acquisition} : \mathbf{R}^{\boldsymbol{X}\times\boldsymbol{U}\times\Lambda} \to \mathbf{R}^{N\times P} \qquad (3)$$

Note that the domain of the operator, *i.e.* the space of light fields, is infinite-dimensional, whereas the co-domain — the space of measures — is finite.

C. LIGHT FIELD COMPRESSED ACQUISITION USING CODED MASKS AND COLOR FILTER ARRAYS

The goal of multi-shot light field acquisition is, given a number of shots N, to recover the light field L. Of course, since the physical description of light fields is infinite-dimensional, we first need to discretize the space of parameters, *i.e.* $\boldsymbol{X}$, $\boldsymbol{U}$ and Λ in order to obtain a computable representation. In the remainder of the article, we use the same notations $\boldsymbol{X}$, $\boldsymbol{U}$ and Λ for the discretized sets, and assume a uniformly spaced discrete parametrization for $\boldsymbol{X}$ and $\boldsymbol{U}$, while Λ is reduced to the common $\{R, G, B\}$ set. By a slight abuse of notation, we will also use $\boldsymbol{X}$ and $\boldsymbol{U}$ to denote the size of those sets, *i.e.* the number of discrete spatial and angular coordinates respectively. The discretized version of Equation 2 is then:

$$I_{n,p} = \sum_{\boldsymbol{X}_p\times\boldsymbol{U}\times\Lambda} L(\boldsymbol{x},\boldsymbol{u},\lambda)\psi_n(\boldsymbol{x},\boldsymbol{u},\lambda) \qquad (4)$$

and the domain of the acquisition operator becomes finite-dimensional. Furthermore, we assume that the pixels $(X_p)_{1\leq p\leq P}$ are squares and correspond to the discretized version of $\boldsymbol{X}$. That is, we assume that $|\boldsymbol{X}| = P$, and that the spatial span of each pixel corresponds to a single discretized element of x: $X_p = \{x_p\}$ for each p, where $(x_p)_{1\leq p\leq P}$ is an enumeration of $\boldsymbol{X}$. Note that one could choose a finer discretization for $\boldsymbol{X}$, in which case the problem would become a joint compressed sensing and super-resolution task. We can thus drop the now redundant notation p and use $\boldsymbol{x}$ instead; Equation 4 becomes:

$$I_n(\boldsymbol{x}) = \sum_{\boldsymbol{U}\times\Lambda} L(\boldsymbol{x},\boldsymbol{u},\lambda)\psi_n(\boldsymbol{x},\boldsymbol{u},\lambda) \qquad (5)$$

The goal of light field compressed multi-shot acquisition is, given a number of shots $N \ll \boldsymbol{U} \times \Lambda$, to recover the whole discretized light field L from the set of measurements. The theory of compressed sensing tells us that the ability to reconstruct the whole signal from a small number of measurements greatly depends on the properties of the sensing operator. As a general principle, the sensing matrix should be as random as possible, so that the measurements be as mutually incoherent as possible [32]. In our case, the acquisition operator is entirely determined by the modulation fields ψ_n.

Our acquisition scheme consists of a *color-coded mask* (CCM) placed at a small distance in front of the photosensor plane, and a *color filter array* (CFA) placed directly on the photosensor. These two elements act as linear filters on the input light field before its projection on the photosensor plane. Figure 1 represents the optical device used for acquisition, with the color-coded mask and the CFA. The CCM is a random mask whereas the CFA is usually a periodic array, traditionally a 2x2 periodic Bayer pattern [42]. In our acquisition framework, multi-shot acquisition can be made possible by allowing the CCM to move on its plane and be

able to be translated by a small amount. Assuming the CCM is characterized by $m : \Xi \times \Lambda \to \mathbf{R}$, where Ξ is the CCM's plane, the corresponding modulation field is derived from elementary geometric considerations:

$$\psi(\boldsymbol{x}, \boldsymbol{u}, \lambda) = m(\underbrace{(1-\gamma)\boldsymbol{x} + \gamma\boldsymbol{u}}_{\boldsymbol{\xi} \in \Xi}, \lambda) \tag{6}$$

where $\gamma = \frac{d(\Xi, \boldsymbol{X})}{d(\boldsymbol{U}, \boldsymbol{X})}$, $d(.,.)$ denoting distance between planes. Equation 6 shows that the sub-aperture slices of the modulation (*i.e.* the restriction of the modulation to a fixed $\boldsymbol{u}$) are all translated versions of one another.

Whereas we could use fixed values for the pixels of the color-coded mask, for the sake of comparison, we instead follow the approach of [15] and [16], in which a different mask is generated each time, by independently drawing the transmittance value $t \in \mathbf{R}^{\Lambda}$ of each of its pixels from a distribution $\mathcal{D}$. Besides, to be able to compare with [15] and [16], we depart from physical realizability in our experiments (unless stated otherwise) by directly drawing a value $t \sim \mathcal{D}$ independently for each element of the modulation field (*i.e.* for each spatial-angular coordinate $(\boldsymbol{x}, \boldsymbol{u})$), instead of computing ψ using Equation (6). However, in order to simulate in a more accurate manner what a real-world set-up would do, we also assess the acquisition scheme using, at test time, a modulation field given by Equation (6) corresponding to a coded mask for which the transmittance of the pixels are drawn from $\mathcal{D}$. We will see in the experimental section (Table 3) that the reconstruction does not suffer from any significant drop in quality when using a modulation field derived from such a physically realizable mask.

Furthermore, it is worth noting that Equation (5) corresponds to an acquisition device that includes pinholes placed on the aperture plane, corresponding to the fact that the real continuous light field is first discretized, and then modulated by way of coded mask and color filter array. Table 3 shows experimental results obtained when simulating the acquisition without pinholes. In that case, a linear interpolation of the available discrete input views is performed to approximate the missing views. The analytical expression of the modulation field given a mask m corresponding to this model is further detailed in Section III-D below.

D. REALISTIC ACQUISITION MODEL WITHOUT PINHOLES

In this section, we derive a more realistic modulation field $\widetilde{\psi}$ from the description of a mask m, assuming a continuous light field that we linearly approximate using the available views. Let us first consider the case of a monochromatic light field for which both the spatial domain and the angular domain are one-dimensional. We assume that the angular domain is discretized into regularly spaced points, so that the angular coordinates where data is available are $u_j = (j - j_0)\Delta u$ for each $j \in [0, \dots, N_u - 1]$, with j_0 such that those coordinates are centered around zero and $\Delta u > 0$ corresponding to the angular resolution. Similarly, the spatial domain is discretized into points located at coordinates $x_i = (i - i_0)\Delta x$ where i_0 is such that the points are centered around zero and $\Delta x > 0$ is the dimension of a sensor pixel. Let $L(x, u)$ denote the continuous light field. We only have access to the discrete data $L_{i,j} := L(x_i, u_j)$. Let m be composed of N_m regularly spaced pixels, such that

$$m(\xi) = \sum_{k=0}^{N_m - 1} m_k \mathbf{1}_{M_k}(\xi) \tag{7}$$

where m_k is the transmittance value on mask pixel $M_k := [\xi_k, \xi_{k+1}]$ with $\xi_k = (k - k_0)\Delta\xi$ and $\Delta\xi > 0$ denoting the dimension of a mask pixel, and $\mathbf{1}_M$ denoting the indicator function of a any subset $M \subset \Xi$.

Since the light fields in our dataset are densely sampled, we assume that L varies smoothly along the coordinates (x, u). Based on this assumption, we approximate the continuous light field at coordinates that are not on the grid by linear interpolation in the angular domain and nearest neighbour interpolation in the spatial domain. That is, $L(.,.) \simeq \widetilde{L}(.,.)$ where:

$$\widetilde{L}(x, u) = \frac{u_{j+1} - u}{\Delta u} L_{i,j} + \frac{u - u_j}{\Delta u} L_{i,j+1} \tag{8}$$

whenever $(x, u) \in [x_i, x_{i+1}] \times [u_j, u_{j+1}]$. The intensity recorded in pixel $X_j = [x_i, x_{i+1}]$ is then given by:

$$I_i = \int_{X_i \times U} m(\xi(x, u))\widetilde{L}(x, u)\mathrm{d}x\mathrm{d}u \tag{9}$$

where $\xi(x, u) = (1 - \gamma)x + \gamma u$. Partitioning the angular domain into pixels $U_j = [u_j, u_{j+1}]$, this equation is rewritten:

$$I_i = \sum_j \int_{X_i \times U_j} m(\xi(x, u))\widetilde{L}(x, u)\mathrm{d}x\mathrm{d}u \tag{10}$$

Injecting Equations (7) and (8) into (10) and rewriting the resulting equation yields

$$I_i = \sum_j \sum_k m_k \cdot (\alpha^{\ell}_{i,j,k} L_{i,j} + \alpha^{r}_{i,j,k} L_{i,j+1}) \tag{11}$$

where

$$\alpha^{\ell}_{i,j,k} = \int_{X_i \times U_j} \mathbf{1}_{M_k}(\xi(x, u)) \frac{u_{j+1} - u}{\Delta u} \mathrm{d}x\mathrm{d}u \tag{12}$$

and

$$\alpha^{r}_{i,j,k} = \int_{X_i \times U_j} \mathbf{1}_{M_k}(\xi(x, u)) \frac{u - u_j}{\Delta u} \mathrm{d}x\mathrm{d}u \tag{13}$$

Further rewriting gives us:

$$I_i = \sum_j \left(\sum_k \beta_{i,j,k} m_k \right) L_{i,j} \tag{14}$$

where $\beta_{i,j,k} = \alpha^{\ell}_{i,j,k} + \alpha^{r}_{i,j-1,k}$, corresponding to a light throughput. Note that the values $\alpha^{\ell}_{i,j,k}$ can be rewritten as

$$\alpha^{\ell}_{i,j,k} = \int_{C_{i,j,k}} \frac{u_{j+1} - u}{\Delta u} \mathrm{d}x\mathrm{d}u \tag{15}$$

where $C_{i,j,k} = (X_i \times U_j) \cap \xi^{-1}(M_k)$ (in which $\xi^{-1}(M_k)$ denotes the pre-image of M_k by ξ), a convex polygon in the spatio-angular domain corresponding to the set of rays intersecting simultaneously the spatial pixel X_i, the mask pixel M_k and the angular pixel U_j. Applying the same rewriting to α^r, one can analytically compute the values for β as a sum of integrals of affine functions over convex polygons.

We now consider light fields that are non-monochromatic and two dimensional in each of the spatial and angular domains and assume that the real continuous light field can be approximated by a bilinear interpolation of the available views. We also resume our notations $\boldsymbol{x}$, $\boldsymbol{u}$, $\boldsymbol{\xi}$ to denote the discrete coordinates corresponding to i, j and k respectively. Equation (14) can be readily extended to obtain the intensity at position $\boldsymbol{x}$ on the sensor for the wavelength λ, before the different wavelengths are modulated by the CFA and integrated over by the sensor, as:

$$I(\boldsymbol{x}, \lambda) = \sum_{\boldsymbol{u}} \left(\sum_{\boldsymbol{\xi}} \boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\xi}) m(\boldsymbol{\xi}, \lambda) \right) L(\boldsymbol{x}, \boldsymbol{u}, \lambda) \quad (16)$$

From Equation (16) it is clear that we can define:

$$\widetilde{\psi}(\boldsymbol{x}, \boldsymbol{u}, \lambda) = \sum_{\boldsymbol{\xi}} \boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\xi}) m(\boldsymbol{\xi}, \lambda) \quad (17)$$

as the modulation field corresponding to our color-coded mask in this discretized view-interpolated framework. Equation (17) is conveniently rewritten as a matrix product:

$$\widetilde{\psi} = \mathbf{B}\mathbf{M} \quad (18)$$

where $\mathbf{B}$ is a $\mathbf{R}^{(X \cdot U) \times \Xi}$ matrix and $\mathbf{M}$ is a $\mathbf{R}^{\Xi \times \Lambda}$ matrix. Since the mask, the sensor and the aperture are all discretized along a separable grid of pixels, *i.e.* all three are the Cartesian product of their respective single-dimensional counterpart with itself, the 2D light throughput $\boldsymbol{\beta}$ is simply computed from its one-dimensional version by:

$$\boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\xi}) = \beta(x, u, \xi_1) \beta(y, v, \xi_2) \quad (19)$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2)$. Note that $\boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\xi})$ equals zero for most combinations of $\boldsymbol{x}$, $\boldsymbol{u}$ and $\boldsymbol{\xi}$, and consequently the corresponding matrix $\mathbf{B}$ is sparse. The modulation field $\widetilde{\psi}$ is therefore efficiently computed as a sparse-dense matrix product. Experimental results using this framework are provided in section V-C.

Please note that Marwah et al [12] successfully built a prototype using a monochromatic coded mask. Manufacturing a color-coded mask may be more challenging than a monochromatic one, nevertheless feasible thanks to recent advances in photolithograpy [43], [44]. In addition, in practice, the modulation ψ cannot be known exactly by analytic means due to imperfections in the manufacturing process. However this problem could be tackled by using a calibration protocol akin to the ones used by [45] and [12].

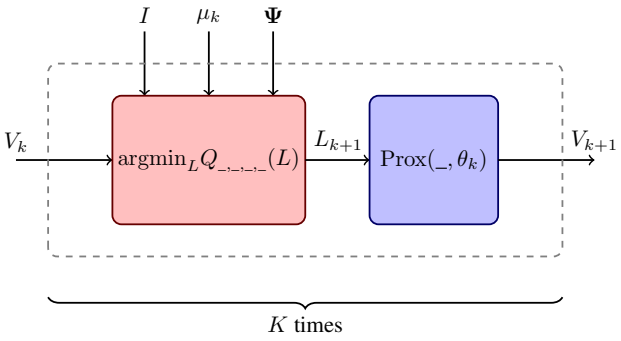


FIGURE 2: **Representation of the unrolled half-quadratic splitting algorithm as a two-layer block.** The first layer corresponds to a (non-trainable) data-term minimization step: it solves a quadratic problem and takes as input an intermediate reconstruction V_k, as well as the measures I, the degradation matrix $\boldsymbol{\Psi}$ and the weight μ_k. The second layer (trainable) projects an intermediate reconstruction onto a sub-manifold of more likely — or natural — light fields and is parameterized by θ_k. The block consisting of these two steps is iterated K times.

IV. ACQUISITION/RECONSTRUCTION ARCHITECTURE

Let $\boldsymbol{\Psi}$ be the sensing matrix corresponding to the acquisition operator defined in the previous section. We have $\boldsymbol{\Psi} \in \mathbf{R}^{\boldsymbol{X}N \times \boldsymbol{X}\boldsymbol{U}\Lambda}$.

Assuming L is the input discretized light field reshaped as a $\boldsymbol{X}\boldsymbol{U}\Lambda$-dimensional vector, the available information is $I = \mathcal{C}(\boldsymbol{\Psi}L) \in \mathbf{R}^{\boldsymbol{X}N}$, where $\mathcal{C}$ is a pixel-wise stochastic corruption process accounting for imperfections in the sensor. A detailed model of $\mathcal{C}$ is given in section VI-A. As a simplification, we can assume $\mathcal{C}$ to be the addition of centered Gaussian noise. In that case, we have:

$$I = \boldsymbol{\Psi}L + \boldsymbol{n} \quad (20)$$

where $\boldsymbol{n}$ is a $\boldsymbol{X}N$-dimensional centered Gaussian random variable with variance σ^2.

A. OPTIMIZATION UNROLLING

The problem of reconstructing the light field L from the measurements I may be formulated in a Bayesian framework, by considering the reconstruction task as a *maximum a posteriori* problem. In its equivalent negative logarithmic formulation, the problem can be expressed as finding L^* such that:

$$L^* = \text{argmin}_L \left(-\log \mathrm{P}(I|L) - \log \mathrm{P}(L) \right) \quad (21)$$

Since the sensing noise is assumed to be centered Gaussian noise, we have:

$$-\log \mathrm{P}(I|L) = \frac{1}{2\sigma^2} \|I - \boldsymbol{\Psi}L\|_2^2 + \text{constant} \quad (22)$$

The reconstruction problem can then be formulated as a regularized least squares problem:

$$L^* = \text{argmin}_L \left(\frac{1}{2} \|I - \boldsymbol{\Psi}L\|_2^2 + J(L) \right) \quad (23)$$

where $J = -\sigma^2 \log \mathrm{P}$.

A popular way to solve this problem is to use the ADMM optimization algorithm [46]. A simpler but effective alternative, especially in the context of algorithm unrolling, is to use the *half-quadratic splitting* (HQS) method [25], [47], [48]. This approach is very similar to ADMM. Just like ADMM, HQS aims at solving unconstrained problems of the form:

$$\mathrm{argmin}_z f(z) + g(z). \tag{24}$$

The problem is first reformulated as a constrained problem, introducing another variable w :

$$\mathrm{argmin}_{z,w} f(z) + g(w) \text{ subject to } z = w. \tag{25}$$

The values of z and w are then iteratively and alternately optimized in an unconstrained framework by relaxing the equality constraint into a half-quadratic penalty term:

$$\begin{cases} z_{k+1} = \mathrm{argmin}_z f(z) + \dfrac{\mu_k}{2}\|z - w_k\|_2^2 \\ w_{k+1} = \mathrm{argmin}_w g(w) + \dfrac{\mu_k}{2}\|w - z_{k+1}\|_2^2 \end{cases} \tag{26}$$

where the $(\mu_k)_{k\geq 0}$ weight the penalty terms. Casting these equations into our regularized least squares problem, and introducing the variables V_k we obtain:

$$\begin{cases} L_{k+1} = \mathrm{argmin}_L \dfrac{1}{2}\|I - \mathbf{\Psi} L\|_2^2 + \dfrac{\mu_k}{2}\|L - V_k\|_2^2 \\ V_{k+1} = \mathrm{argmin}_V J(V) + \dfrac{\mu_k}{2}\|V - L_{k+1}\|_2^2 \end{cases} \tag{27}$$

The update rule for L clearly corresponds to the resolution of a quadratic problem, whereas the update rule for V can be interpreted as applying a proximal operator Prox. We further rewrite in a more synthetic form:

$$\begin{cases} L_{k+1} = \mathrm{argmin}_L Q_{\mu_k, I, \mathbf{\Psi}, V_k}(L) \\ V_{k+1} = \mathrm{Prox}_k(L_{k+1}) \end{cases} \tag{28}$$

Now, while it is clear that the update rule for L can be easily applied, at least theoretically, given I, $\mathbf{\Psi}$ and V, the proximal operator actually depends on the underlying probability distribution P on light fields. Traditional methods have been somewhat successful at recovering signals from compressed measurements using hand-crafted choices of P, yet recent achievements in low-level image processing suggest that it is immensely beneficial to learn these priors from data. *Algorithm unrolling* in deep learning consists in using Equation 28 to design a deep learning architecture. Specifically, one usually fixes a number of iterations K and then, instead of considering the Prox_k as fixed functions, one replaces them by K trainable parameterized functions $\mathrm{Prox}(\_, \theta_k)$, typically neural networks. The general architecture for the unrolled half-quadratic splitting algorithm is depicted in Figure 2.

To our knowledge, the minimization of the quadratic terms is not solved in an exact manner in methods of the literature using unrolled HQS. Instead, some authors, for instance [25], replace the exact resolution by a single step of gradient descent on the quadratic function. The authors in [47] argue that the problem cannot be resolved in closed form, due to the complexity and size of the sensing matrix $\mathbf{\Psi}$. While this may be true for some inverse problems, we show that in our compressed sensing framework the quadratic problem can actually be solved in closed form efficiently.

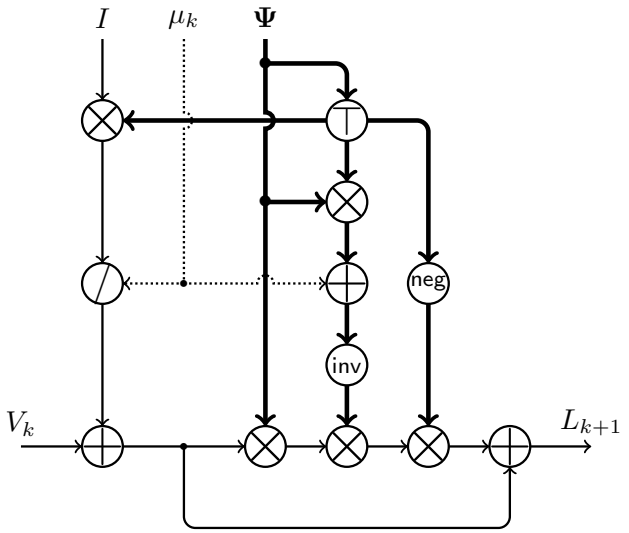


FIGURE 3: **Network representation of the data-projection layer by efficient closed-form resolution of the quadratic problem.** Multiplication nodes represent either matrix-matrix multiplication or matrix-vector multiplications. Thick lines indicate matrices, normal lines indicate vectors and dotted lines indicate scalars. Note that μ_k is applied element-wise in the division node, but only on the diagonal in the matrix-scalar addition node. The nodes neg, inv and $\top$ indicate multiplication by -1, matrix inversion and matrix transposition respectively.

B. CLOSED-FORM SOLUTION OF THE DATA-FIDELITY TERM MINIMIZATION

A single step gradient descent is often used in the literature, as in [25], to minimize the data term of unrolled optimization methods, since the inversion of the corresponding degradation matrices is in general computationally untractable. However, this only gives an approximation of the optimal solution to the minimization of the data term. In our proposed scheme, we show that the structure of our sensing matrix is such that this matrix can be inverted in a very efficient manner using the Sherman-Morrison-Woodbury identity.

Taking the gradient of the quadratic form in the L-update of Equation 27 (with respect to L), we get:

$$\nabla_L Q = \mathbf{\Psi}^\top(\mathbf{\Psi} L - I) + \mu_k(L - V_k) \tag{29}$$

The single-step gradient descent update rule for L would be:

$$\begin{aligned} L_{k+1}^{\mathrm{SS}} &= L_k - \delta_k \nabla_L Q(L_k) \\ &= ((1 - \delta_k \mu_k)\mathbf{I} - \delta_k \mathbf{\Psi}^\top \mathbf{\Psi}) L_k \\ &\quad + \delta_k \mathbf{\Psi}^\top I + \delta_k \mu_k V_k \end{aligned} \tag{30}$$

where $\delta_k > 0$ is the gradient descent rate.
The closed-form solution of the equation $\nabla_L Q = 0$ is given by:

$$L_{k+1}^{\mathrm{CF}} = L^* = (\mathbf{\Psi}^\top \mathbf{\Psi} + \mu_k \mathbf{I})^{-1} (\mathbf{\Psi}^\top I + \mu_k V_k) \qquad (31)$$

Therefore we see that, *a priori*, in order to compute the closed-form solution L^*, one has to invert the $\boldsymbol{XU\Lambda} \times \boldsymbol{XU\Lambda}$ regularized covariance matrix $\mathbf{\Psi}^\top \mathbf{\Psi} + \mu_k \mathbf{I}$. However, recall from section III-C that our sensing matrix $\mathbf{\Psi}$ is defined by Equation 5. This equation shows that the multiplexing of information acquired by the sensor is actually performed "spatial pixel-wise". This means that our compressed acquisition sensing matrix is actually block-diagonal. We can write:

$$\begin{pmatrix} I_1(\boldsymbol{x}_1) \\ I_2(\boldsymbol{x}_1) \\ \vdots \\ I_N(x_1) \\ \vdots \\ I_1(\boldsymbol{x}_{|\boldsymbol{X}|}) \\ I_2(\boldsymbol{x}_{|\boldsymbol{X}|}) \\ \vdots \\ I_N(\boldsymbol{x}_{|\boldsymbol{X}|}) \end{pmatrix} = \begin{pmatrix} \mathbf{\Psi}_1 & 0 & \dots & 0 \\ 0 & \mathbf{\Psi}_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \mathbf{\Psi}_{|\boldsymbol{X}|} \end{pmatrix} \begin{pmatrix} L(\boldsymbol{x}_1, \boldsymbol{u}_1, \lambda_1) \\ L(\boldsymbol{x}_1, \boldsymbol{u}_1, \lambda_2) \\ \vdots \\ L(\boldsymbol{x}_1, \boldsymbol{u}_{|\boldsymbol{U}|}, \lambda_{|\Lambda|}) \\ \vdots \\ L(\boldsymbol{x}_{|\boldsymbol{X}|}, \boldsymbol{u}_1, \lambda_1) \\ L(\boldsymbol{x}_{|\boldsymbol{X}|}, \boldsymbol{u}_1, \lambda_2) \\ \vdots \\ L(\boldsymbol{x}_{|\boldsymbol{X}|}, \boldsymbol{u}_{|\boldsymbol{U}|}, \lambda_{|\Lambda|}) \end{pmatrix} \qquad (32)$$

where each $\mathbf{\Psi}_{\boldsymbol{x}}$ is a $N \times \boldsymbol{U\Lambda}$ matrix characterizing the multiplexing happening on pixel $\boldsymbol{x}$. Thus the inversion of the matrix can be performed efficiently block-wise — *i.e.* pixel-wise — by computing, possibly in parallel, the matrices $(\mathbf{\Psi}_{\boldsymbol{x}}{}^\top \mathbf{\Psi}_{\boldsymbol{x}} + \mu_k \mathbf{I})^{-1}$. As a notational aside, we drop the subscript $\boldsymbol{x}$ in the remainder, and regard the block-diagonal matrix $\mathbf{\Psi}$ as a $\boldsymbol{X} \times N \times \boldsymbol{U\Lambda}$ tensor; all operations on matrices are regarded as being performed pixel-wise.

Nonetheless, these matrices are still of size $\boldsymbol{U\Lambda}$, so a direct computation, though tractable, remains somewhat costly (as $\boldsymbol{U\Lambda}$ is usually in the order of 10^2). This cost can be further alleviated by using the Sherman-Morrison-Woodbury identity [49]:

$$(\mathbf{\Psi}^\top \mathbf{\Psi} + \mu_k \mathbf{I})^{-1} = \mu_k^{-1} (\mathbf{I} - \mathbf{\Psi}^\top (\mathbf{G} + \mu_k \mathbf{I})^{-1} \mathbf{\Psi}) \qquad (33)$$

where $\mathbf{G}$ is the Gram matrix $\mathbf{\Psi}\mathbf{\Psi}^\top$. Note how the matrices to invert $(\mathbf{G} + \mu_k \mathbf{I})^{-1}$ are now of size $N \ll \boldsymbol{U\Lambda}$.
By denoting $Z_k = \mu_k^{-1} \mathbf{\Psi}^\top I + V_k$, we have:

$$L_{k+1} = Z_k - \mathbf{\Psi}^\top (\mathbf{G} + \mu_k \mathbf{I})^{-1} \mathbf{\Psi} Z_k \qquad (34)$$

and the product on the right-hand side is efficiently computed by matrix-vector multiplications, once the $\mathbf{G} + \mu_k \mathbf{I}$ have been inverted. All these multiplications can be carried out pixel-wise in parallel efficiently. Figure 3 gives a network representation of the L-update step. It is worth noticing that this architecture makes it straightforward and inexpensive to learn not only the parameter μ_k, but also the sensing matrix $\mathbf{\Psi}$. This point is further studied in section VI-B.

C. LEARNED PROXIMAL OPERATOR

While the structure of the function corresponding to the proximal operator is the same for each unrolled iteration, the weights between them are not shared. This is customary

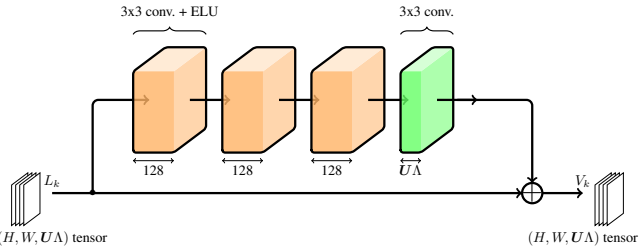


FIGURE 4: **Convolutional network corresponding to one learned proximal operator.** The residual branch consists in 4 stacked convolutions and element-wise non-linear activation functions. Each convolution in the residual branch has 128 filters, except the last one which has $\boldsymbol{U\Lambda}$ to recover the full dimension of the signal.

practice and allows for a greater expressive power of the overall architecture. We use a residual structure for our proximal operator, with a single skip-connection (see [50]), *i.e.* we have

$$\mathrm{Prox}_k(L) = L + \mathrm{Res}(L, \theta_k) \qquad (35)$$

where Res is a function chosen to be a simple 2-dimensional convolutional neural network (CNN) with ELU non-linear activation [51]. The input light field L of shape $(\boldsymbol{X}, \boldsymbol{U}, \Lambda)$ is first reshaped into a $(\boldsymbol{X}, \boldsymbol{U\Lambda}) = (X, Y, \boldsymbol{U\Lambda})$ tensor, which is processed as a 2D image with $\boldsymbol{U\Lambda}$ channels by the CNN. The output of the CNN is then reshaped back to its original shape. Each convolution except the last one in the residual branch consists in a 3×3 kernel with 128 filters. The last convolution uses a 3×3 kernel with $\boldsymbol{U\Lambda}$ filters to match the dimensions of the light field. The last convolution block does not have any subsequent non-linear activation, in order to maintain a full range of values for the residual. In our experiments we use a stack of 4 convolutions. The architecture of the proximal network is depicted in Figure 4.

D. OVERALL ARCHITECTURE

The overall reconstruction architecture is composed of the elements described in IV-A, IV-B and IV-C. The initial reconstruction V_0 is set to 0. We also add a final clipping layer, so that the output of the reconstruction architecture is $\hat{L} := \max(0, \min(1, V_K))$, to ensure that the intensities of the reconstructed light field lie between 0 and 1. In our experiments we use a number of unrolled iterations $K = 12$. For each iteration, the trainable proximal operator has $4.7 \cdot 10^5$ parameters, in addition to the learned μ_k parameter. Because the weights are not shared between the different layers, the reconstruction network has a total of $5.6 \cdot 10^6$ parameters. In addition to the reconstruction system which is purely computational, some physical parameters, namely the pattern of the CFA and the CCM, are learnable in our framework, since it is possible to physically produce an optical device with an arbitrary CFA and CCM. While [16] reported the benefits of learning the CFA and the distribution of the CCM, we found no improvement over using a fixed CFA and fixed

CCM distribution in the noiseless case, with the proposed architecture. We instead used randomly generated CCMs in which the transmittance values of the pixels are drawn from a distribution $\mathcal{D} = \mathcal{U}(0, 1)$, and a fixed 2×2 Bayer CFA (green-blue-red-green). The end-to-end acquisition-reconstruction process is summarized in Figure 5. We hypothesise that the inductive bias towards inverse problem solving in the architecture, enforced by the closed-form data-term minimization layer, makes the network less sensitive to the specifics of the degradation matrix, at least in the absence of noise.

Nonetheless, learning the CFA and distribution of the CCM proved to be greatly beneficial in the noisy case, as detailed in section VI-C.

V. EXPERIMENTS

A. TRAINING DETAILS

We defined and trained our model using TensorFlow 2 [1]. We used the Stanford Lytro light field archive [52] as our training dataset, and used the Linköping University Lytro dataset [2] as a validation set (14 light fields). We performed our experiments using light fields with an angular resolution of 5×5 views. For practical purpose, and since the overall architecture is convolutional, we can train the model using light field patches of a smaller resolution, and use full-size light fields at test time. We chose a patch resolution of 64×64 pixels; patches are extracted from the full-size light fields using a stride of 32 pixels. No data augmentation is applied. We used the ℓ^1 norm for our regression as usual for this kind of task. The whole network was trained using the ADAM [53] optimizer, using hyper-parameters $(\beta_1, \beta_2, \epsilon) := (0.9, 0.98, 10^{-5})$ and gradient clipping to maintain the values within the range $[-1, 1]$. We empirically found that these values prevented explosion of the loss that could happen otherwise during training, especially as the number of unrolled iterations in the model increases. The learning rate was set to 10^{-3} for the first $5 \cdot 10^5$ steps, and then decreased to 10^{-4} until the end of training. Each model was trained for $8 \cdot 10^5$ steps using a batch size of 16. This corresponds to 180 epochs, each epoch containing about $7 \cdot 10^4$ sample patches extracted from 476 light fields. For stability purposes, we additionally clip the values of the μ_ks at each optimization step so that they remain greater than 10^{-2}.

B. RESULTS

We first trained and evaluated our architecture with a number of shots $N \in \{1, 2, 3\}$ in a noiseless framework (that is, where the corruption operation $\mathcal{C}$ is the identity). In this first experiment, we use a modulation field in which the transmittance of each of its elements is independently drawn according to a uniform distribution. Each model for a given number of shots was trained separately, and all follow the same training schedule as detailed in section V-A.

Table 1 provides a quantitative comparison of different models, in terms of peak signal-to-noise ratio (PSNR), for a set of light fields used in all the methods referenced in the table. In the sequel, we refer this set of light fields as the "base test set". Table 1 also compares the proposed algorithm with the closed-form solution of the data term minimization against the single-step unrolled HQS architecture, which uses the same architecture for the learned proximal operator, but performs a single gradient descent step of data-term minimization as prescribed by Equation 30 instead of solving it in closed form. These models have an additional trainable parameter δ_k for each layer, which we all initialize to 0.05, a hyper-parameter that was tuned to perform reasonably well for all three numbers of shots. We see that using the closed-form solution improves the reconstruction quality by about 0.6 dB, regardless of the number of shots used, and systematically outperform the single-step approach except in one case (see Table 2). We also include in our comparison models from [15], [16] and [14], for which the acquisition scheme is identical to ours.

We additionally compare with [25], which uses a different compressed acquisition scheme based on coded apertures. Since their method reconstructs a light field from a set of RGB measurements (instead of monochrome ones in our method), for a fair comparison we compare their method using a single RGB measurement to our method using 3 monochrome measurements. Note that we used the multiple-channel version of [25] where all three color channels are reconstructed jointly from the RGB measurements, thus allowing the exploitation of cross-channel correlations. The model was retrained using the code provided by the authors. A more extended comparative analysis with the model of [25] is given in the next sections which studies different aspects of our camera design. Tables 3 gives a comparison of the various methods both in terms of average PSNR and average EPI-SSIM (an angular consistency metric that averages the SSIM of all epipolar images) for several datasets.

Figures 6 and 7 visually compare the reconstruction of the central view with different methods, and also show the reconstruction error. Figure 8 shows epipolar plane images (EPI) of the reconstructed views. We can see that our method reconstructs the parallax correctly, an essential aspect of light fields. More material assessing the visual quality of the different methods can be found at `clim.inria.fr/research/DeepUnrolling-CSLF`.

C. COMPARISON OF DIFFERENT ACQUISITION SCHEMES

We compare our main acquisition scheme that uses a modulation field in which each element is independently drawn from a uniform distribution (thus not physically corresponding to a real coded mask) to a scheme that uses a modulation field corresponding to a real coded mask in the presence of pinholes on the aperture, as given by Equation (6). For this comparison, we did not retrain the whole model,

[1] The code is accessible at `github.com/gleguludec/deepulfcam` or *here* through ftp.

[2] `computergraphics.on.liu.se/hdrlf`

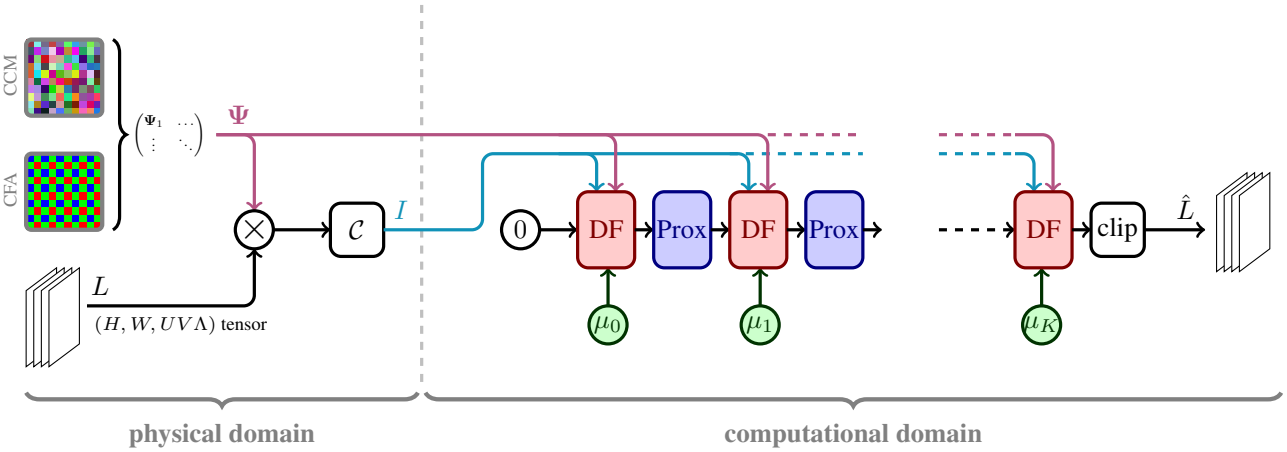


FIGURE 5: **End-to-end architecture.** The left-hand part models the physical operations, comprising the modulation by CCM and CFA and photo-sensing (multiplication by the sensing matrix) and the stochastic corruption process $\mathcal{C}$ — this part is only present in the noisy framework. The right-hand part represents the deep unrolled algorithm: the reconstruction is first initialized with value 0, and then the following are applied: data-term minimization step, or data-fidelity layer (DF) — whose architecture is detailed in Figure 3, intertwined with the application of proximal operators (Prox) — whose architecture is detailed in Figure 4. A final clipping operation is applied to ensure that the values of the reconstructed intensities lie in a valid range.

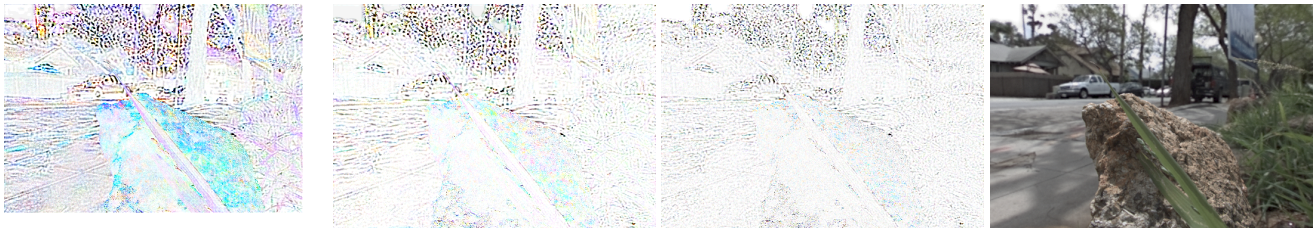
Number of shots	1					2			3			
Method	TV-Dict [14]	[15]	[16]	SS-HQS	CF-HQS	TV-Dict [14]	SS-HQS	CF-HQS	TV-Dict [14]	[25]	SS-HQS	CF-HQS
Buttercup	29.64	29.98	32.41	34.63	**34.91**	32.67	36.14	**36.55**	33.76	36.29	37.21	**37.72**
Cars	27.12	29.88	31.22	32.53	**33.25**	31.98	34.82	**35.51**	33.37	35.53	36.02	**36.62**
Orchids	28.41	30.99	32.89	33.30	**34.15**	33.00	36.16	**37.22**	34.16	36.07	36.69	**37.75**
Rock	27.15	30.11	31.22	33.99	**34.97**	33.23	36.89	**37.66**	34.69	37.34	38.12	**38.96**
Seahorse	29.92	32.36	33.61	34.42	**35.39**	33.16	37.40	**38.04**	35.26	37.95	38.59	**39.14**
Tulips	40.83	38.26	42.89	46.53	**46.87**	43.03	48.07	**48.35**	44.47	45.75	48.74	**49.03**
White rose	28.23	31.84	32.96	33.57	**34.49**	33.29	36.43	**37.00**	34.71	37.00	37.55	**38.09**
Average	30.19	31.92	33.89	35.56	**36.29**	34.34	37.70	**38.54**	35.77	37.99	39.07	**39.62**

TABLE 1: **PSNR comparison of the different methods (in dB), with a modulation field in which the transmittance of each element is independently drawn according to a uniform distribution.** CF-HQS stands for our "closed-form HQS" architecture in which the data-term minimization is performed using Equation (34), SS-HQS stands for the "single-step HQS" architecture which uses Equation (30) to perform the data-term minimization step. [14] is an iterative dictionary-based approach, while [15], [16] and [25] are based on deep neural networks. Note that [15] and [16] work only with a single-shot acquisition, while on the contrary [25] uses RGB measurements, thus being comparable only to our monochromatic 3-shot acquisition scheme.

but merely changed the structure of the modulation field at test time. The modulation field for each sub-aperture view was obtained by translating the 2D mask by 8 pixels per view, which corresponds to a parameter $\gamma \simeq 0.1$ if we make the reasonable assumptions that the sensor is about 400 pixels wide and that the 5×5 pinholes span an aperture which has approximately the same size as the sensor. Table 3 shows no perceptible drop in performance when using such physically accurate modulation field instead of the one used for training. We hypothesize that this is because the modulation fields used during training are general enough to allow the network to perform well in the specific case in which the modulation field derives from a coded mask. In addition, even though the sub-aperture views of the CCM are translated versions of each other, the norm of the translation is in practice larger than the receptive field of the reconstruction algorithm which may have the effect that the sub-aperture views of the CCM locally look like they have been drawn independently from one another.

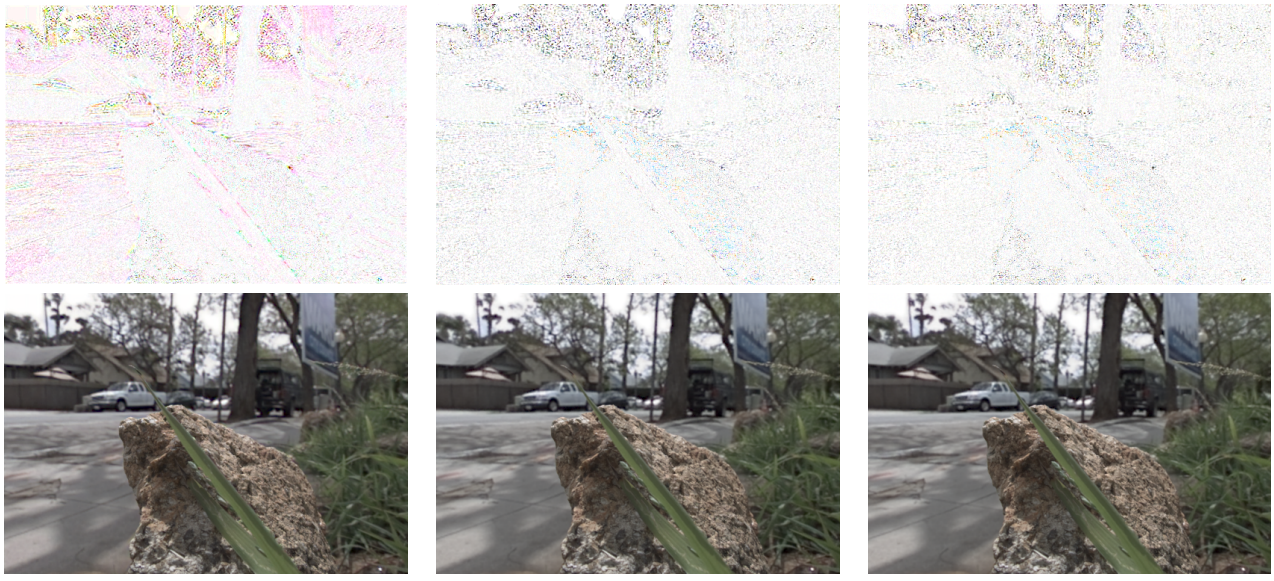
In addition to these pinholes-based schemes, we give experimental results when using the scheme without pinholes with bilinear interpolation of the views described in section III-D. For this scheme, we provide the results for one experiment using a new mask randomly generating by drawing the transmittance value of each pixel of the mask from $\mathcal{U}(0, 1)$. We performed our experiments with a relative distance parameter $\gamma = 0.1$, so as to compare with the pinhole-based model. We show in Table 3 that this model performs well, even outperforms the pinhole-based model on half of the test datasets.

Finally, we compare our coded mask acquisition scheme to a monochromatic coded aperture scheme, which is the one



(a) *Nabati et al. [15]* **(b)** *Le Guludec et al. [16]* **(c)** *CF-HQS (ours)* **(d)** *Ground truth*

FIGURE 6: **Reconstruction error for the central view of the light field *Rock*, with a single-shot acquisition.** For a better visual readability, we amplified the error by a factor 10. Lighter means lower error, darker means higher error. Figure (a) is cropped in the bottom-right corner (reconstruction provided by the authors).



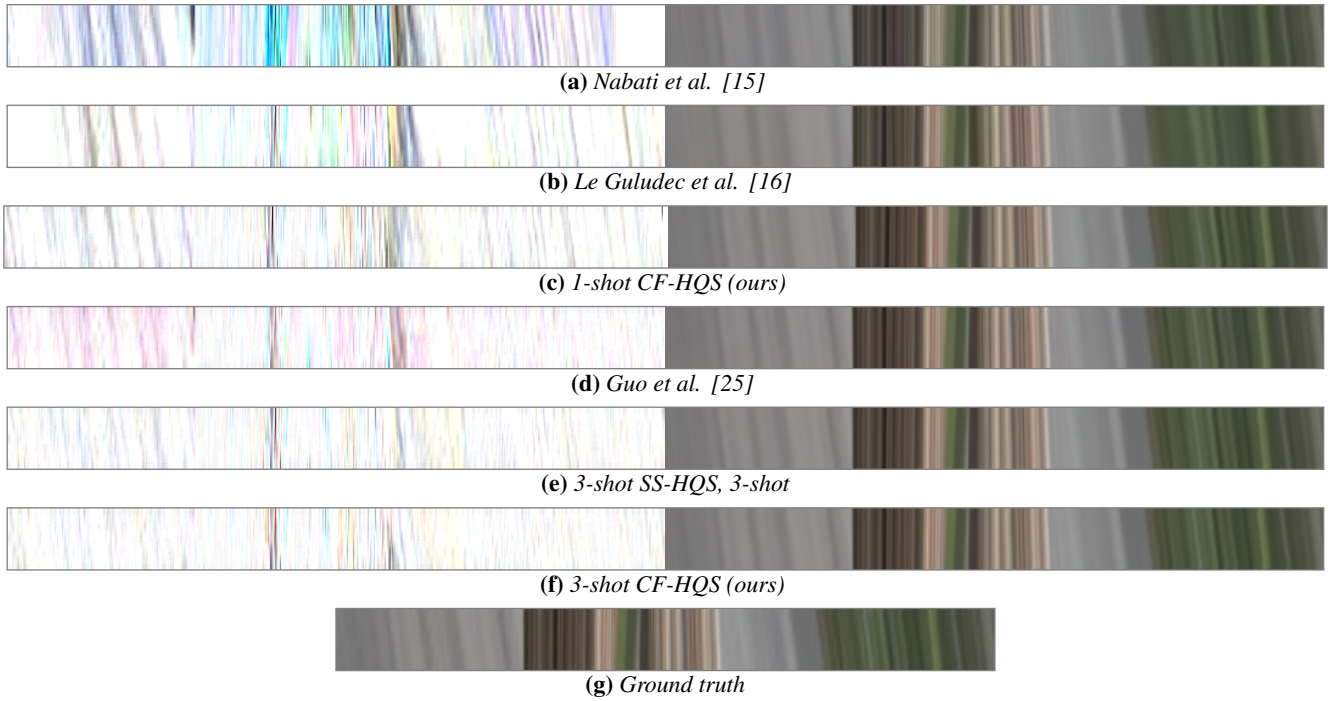
(a) *Guo et al. [25]* **(b)** *SS-HQS* **(c)** *CF-HQS (ours)*

FIGURE 7: **Reconstruction and reconstruction error for the central view of the light field *Rock*, with a 3-shot acquisition.** For a better visual readability, we amplified the error by a factor 10. Lighter means lower error, darker means higher error.

used in [25], [31] and [30]. In these designs, the mask is placed at the aperture plane, whereas in our design it is placed close to the sensor. Furthermore, note that all the experiments presented in Table 3 use a Bayer color filter array in addition to the mask, thereby recording only monochromatic measurements but allowing multiplexing in the spectral domain, while [25] and [30] use true RGB sensors, thereby having three times more channels for each shot.

Following the scheme used in our other experiments, each pixel of the coded aperture is randomly drawn from a uniform distribution for each light field sample. Table 3 shows that using a monochromatic coded aperture yields worse results comparing to using a color-coded mask. We hypothesise that its placing close to the sensor actually allows the rays coming from different angles to be multiplexed in a way that is dependent on the spatial position of the incident pixel, thereby increasing the randomness, hence the possibility for the various measurements to be mutually incoherent, which is known to be a crucial property of degradation matrices as explained by [32].

It is also worth noting that in all of our experiments, we do not learn a fixed degradation operator, but instead sample a new degradation operator each time a light field sample is processed. As a consequence, for a given acquisition scheme, our network is not tied to a particular realisation of the degradation operator, which makes our approach more robust to the specifics of the degradation operator (for the purpose of physically realizing our acquisition device, it suffices to fix the degradation to one particular sample from the distribution). This sets us apart from the framework of [25], [31] and [30]. By learning the degradation operator (instead of sampling a new one each time a light field is processed), [25] and [31] were able to relax the constraint that the weights of the degradation operator should be shared between the various layers, so as to increase the expressive power of their network, thus learning different *virtual* degradation matrices at each step of the unrolled algorithm. Note that in the case of [25], [31] and [30], the degradation is obtained by angular



(a) *Nabati et al. [15]*

(b) *Le Guludec et al. [16]*

(c) *1-shot CF-HQS (ours)*

(d) *Guo et al. [25]*

(e) *3-shot SS-HQS, 3-shot*

(f) *3-shot CF-HQS (ours)*

(g) *Ground truth*

FIGURE 8: **Reconstructed EPI (right-hand side) and reconstruction error of EPI (left-hand side) of the light field** *Rock***.** Figures (a) to (c) are comparable as they correspond to single-shot acquisition methods, while figures (d) to (f) correspond to multi-shot acquisition schemes. For a better visual readability, we amplified the error by a factor 10. Lighter means lower error, darker means higher error.

multiplexing of the light field and is thus spatially invariant, allowing the authors of [25] and [31] to implement their relaxed degradation operators at each unrolled iteration as angular convolutions and deconvolutions. In contrast, the degradation operator resulting from a coded mask acquisition scheme is not invariant to translation in the spatial domain and therefore cannot be implemented using angular convolution and deconvolution layers. Besides, the stochasticity of the degradation prevents us from directly learning additional weights corresponding to the relaxed degradation operators. However, relaxing the sharing condition comes at the cost of interpretability. Indeed, this departs from the original unrolled optimization algorithm and prevents us from considering the output of these layers as a projection on the space of measurement-consistent signals.

D. IMPACT OF THE STRUCTURE OF THE REGULARIZER

Table 2 gives averaged PSNR and EPI-SSIM results on several datasets obtained with a regularizer using 2D convolutions in comparison with an architecture using a regularizer consisting of a stack of spatio-angular separable 4D convolutions (as used in [25] and [31]), and also compares with [25] and [31]. Using 4D convolution-based regularizers instead of the 2D ones generally yield slightly better results, however at the cost of a greatly increased computation time. More precisely, 4D separable convolutions have sensibly fewer parameters than traditional 2D convolutions performing operations on flattened light fields. The time needed (during training) to process one sample is one order of magnitude longer than that of 2D convolutions, but the learning converges faster, which leaves the overall time required for training approximately unchanged. However, the extra computational time for each sample remains an issue at test time: when tested on an Nvidia Quadro RTX 8000, the reconstruction of a single light field takes about 0.3 seconds with the 2D convolutional regularizer, whereas it takes approximately 4 seconds with the 4D separable approach.

E. IMPACT OF THE NUMBER OF UNROLLED ITERATIONS

We studied the impact of increasing the number of unrolled iterations K. While increasing the number of iterations could in theory improve the final reconstruction quality, we observe a quick saturation. Figure 9 shows the effect of the number of unrolled iterations on the quality of reconstruction. Increasing K from 8 to 12 improves the PSNR by 0.78 dB, while further increasing K from 12 to 16 yields an average improvement of only 0.02 dB.

VI. EXPERIMENTS IN A NOISY SETTING

A. CORRUPTION MODEL

While the results presented in V-B were all conducted in a noiseless framework, we additionally evaluated the robustness of our approach in a noisy framework. To that end, we applied a corruption process to the monochromatic images

Number of shots	Method	Base test set	Inria Illum [54]	Inria synthetic [55]	HCI (new) [56]
1	SS-HQS	35.57 / 0.9603	27.75 / 0.7672	28.63 / 0.8210	30.36 / 0.7251
	SS-HQS 4D	35.90 / 0.9609	**28.41 / 0.7791**	**29.48** / 0.8337	**30.96 / 0.7362**
	CF-HQS	**36.34 / 0.9641**	28.02 / 0.7683	29.29 / 0.8347	30.73 / 0.7274
	CF-HQS 4D	35.51 / 0.9590	27.98 / 0.7722	29.27 / **0.8382**	30.58 / 0.7345
3	SS-HQS	39.12 / 0.9777	29.82 / 0.8134	32.48 / 0.8940	33.22 / 0.7794
	SS-HQS 4D	39.57 / 0.9794	29.56 / 0.8097	32.96 / 0.9033	33.22 / 0.7813
	CF-HQS	39.62 / 0,9795	**30.04** / 0,8153	32.81 / 0,8982	33.46 / 0,7843
	CF-HQS 4D	**40.25 / 0.9815**	30.02 / **0.8177**	**33.96 / 0.9134**	**33.84 / 0.7873**
$1 \times$ RGB	Guo *et al.* [25]	37.99 / 0.9767	29.8 / 0.8149	30.69 / 0.8686	32.54 / 0.7824
$2 \times$ RGB	Guo *et al.* [25]	**40.19 / 0.9829**	30.79 / 0.8447	31.40 / 0.8925	33.26 / 0.8097
	Guo *et al.* [31]	40.00 / 0.9823	**33.43 / 0.8839**	**32.7 / 0.8985**	**35.32 / 0.8578**

TABLE 2: **Averaged PSNR and EPI-SSIM measures obtained with different schemes:** our unrolled HQS method using single step gradient descent (SS-HQS) or the closed-form expression (CF-HQS), with a regularizer using 2D or 4D convolutions. We compare the results with 1 and 3 shots against the methods in [25] and [31]. Note that 1 shot with the acquisition in RGB (as in [25] and [31]) is equivalent to 3 shots with our acquisition scheme, in terms of number of input measurements. The base test set corresponds to the light fields used in Table 1.

Acquisition matrix	Base test set	Inria Illum [54]	Inria synthetic [55]	HCI (new) [56]
CCM-independent	36.34 / 0.9641	**28.02 / 0.7683**	29.29 / 0.8347	**30.73** / 0.7274
CCM-pinhole	36.23 / 0.9642	27.98 / 0.7679	29.17 / 0.8350	30.70 / **0.7276**
CCM-views-interpolation	**36.76 / 0.9686**	27.85 / 0.7657	**29.31 / 0.8411**	29.57 / 0.6926
MCA	31.83 / 0.9361	26.52 / 0.7484	24.77 / 0.7528	27.17 / 0.6920

TABLE 3: **Averaged PSNR and EPI-SSIM measures obtained with different acquisition models, with one shot, and different datasets.** *CCM-independent* is our base model, for which the modulation field is generated by randomly drawing the transmittance value of each of its elements from a uniform distribution. *CCM-pinhole* corresponds to the case where the modulation field is related to a uniformly generated coded mask by Equation (6). *CCM-views-interpolation* corresponds to the case where the modulation field is related to a uniformly generated coded mask by Equation (18). Finally, *MCA* stands for monochromatic coded aperture.

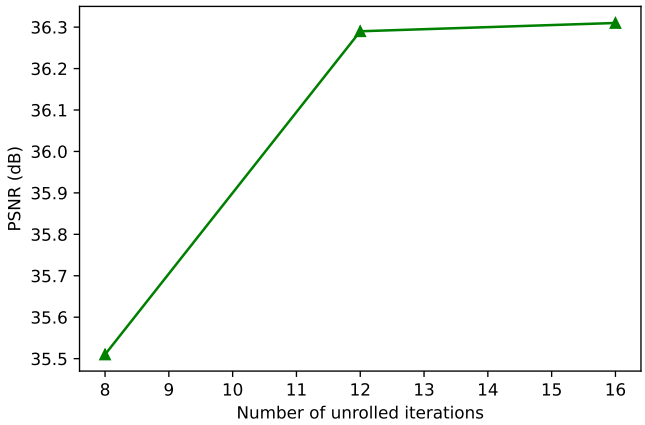


FIGURE 9: **Impact of the number of unrolled iterations on the reconstruction quality.** Average reconstruction quality in PSNR (dB) on the test set, for a number of shots $N = 1$.

captured by the photosensor. Following [14] and [16], we include in our noise model three sources of corruption: *shot noise*, caused by the quantized nature of the light reaching the sensor; *readout noise*, which originates from the spontaneous emission of electrons by the photoelectric device and *quantization noise* arising during the analog-digital conversion of electrons into digital units. The corrupted pixel-wise intensity recorded on the photo-sensor is given by:

$$\mathcal{C}(I) = \frac{1}{gB}\text{int}\left(\frac{B}{c}\text{clip}_{[0,c]}(\mathbf{p}(gcI) + \mathbf{n}_{\text{rd}})\right) \tag{36}$$

where $\text{int}(.)$ denotes rounding, $\text{clip}_{[0,c]}(.) = \max(\min(., c), 0)$, $B = 2^b - 1$, 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 an intensity gain factor proportional to the ISO gain of the pixel and the exposure time; $\mathbf{p}(\alpha)$ is a random variable following a Poisson distribution $\mathcal{P}(\alpha)$ and $\mathbf{n}_{\text{rd}} \sim \mathcal{N}(0, \sigma_{\text{rd}}^2)$. However, the discrete nature of some of the sources of randomness makes it difficult to integrate them into a differentiable gradient-based learning approach. For this reason, we approximate the corruption

process using continuous distributions. We thus substitute the Poisson distribution $\mathcal{P}(\alpha)$ modeling shot noise with a Gaussian distribution $\mathcal{N}(\alpha, \alpha)$, and the rounding $\mathrm{int}(.)$ with an additive uniform noise $\mathbf{u} \sim \mathcal{U}(-\frac{1}{2}, \frac{1}{2})$. Since the shot and readout noises are independent from each other, their sum follows the Gaussian distribution $\mathcal{N}(gcI, gcI + \sigma_{\mathrm{rd}}^2)$.

Our actual model for the corrupted measurements therefore becomes:

$$\mathcal{C}(I) = \frac{1}{gc}\mathrm{clip}_{[0,c]}\left(gcI + \sqrt{gcI + \sigma_{\mathrm{rd}}^2}\mathbf{n}\right) + \frac{1}{gN}\mathbf{u} \quad (37)$$

where $\mathbf{n}$ follows the standard normal distribution $\mathcal{N}(0, 1)$.

B. LEARNING THE CCM AND CFA

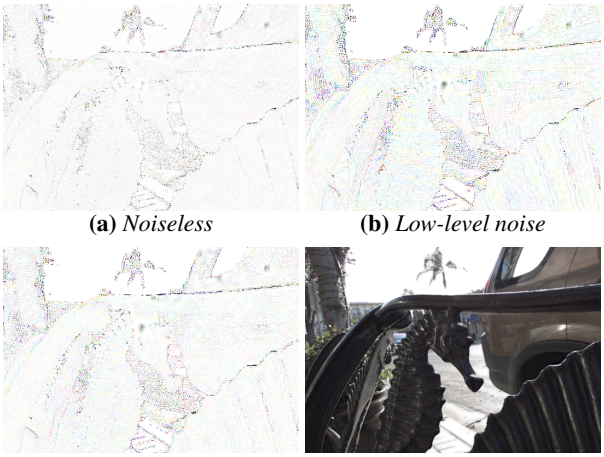
While using a Bayer CFA and uniform CCM may be an efficient way to multiplex spectral information in a noiseless framework, their low transmittance ($\frac{1}{3}$ and $\frac{1}{2}$ respectively, yielding an even lower transmittance of $\frac{1}{6}$ for the global device) make them ill-suited in a noisy framework, as the signal-to-noise ratio is greatly damaged. This makes it desirable to actually learn the CFA and CCM, along with the weights of the reconstruction network. For the CFA we simply learn a periodic 4×4 pattern, which yields 48 additional parameters to the overall model. These weights are initialized following a uniform distribution between 0 and 1. Following the considerations in section III-C and the approach of [16], we learn a distribution of the CCM by learning a distribution $\mathcal{D}$ of the transmittance on the pixels of the CCM. This 3-dimensional distribution is learned using a simple multi-layer perceptron with output dimension 3 that we feed with random noise. More precisely, we sample a $\boldsymbol{XU} \times d$-dimensional standard Gaussian random variable s, and set $\psi(\boldsymbol{x}, \boldsymbol{u}, .) := \mathrm{MLP}_\Theta(s)$ — *i.e.* $\mathcal{D}$ is the direct image of $\mathcal{N}(0, \mathbf{I})$ under MLP_Θ. The parameters Θ of the MLP are then learned using standard back-propagation. In our experiments, we use a stack of three dense layers (*i.e.* affine transformations) with 32 hidden layers, interleaved with ReLU non-linearities, and with a final logistic non-linearity to ensure that the output transmittance lie in the range $(0, 1)$. This yields 1443 additional parameters, a negligible amount compared to the number of parameters of the whole network.

In addition, we use an entropy-based regularization scheme on the pixels transmittance values distribution, as it was shown in [16] to be an effective way to prevent the learned distribution from falling into a poor local minimum.

Please note also that, even though it might be difficult to create a pixel on the mask with any arbitrary transmittance, the constraints imposed on the pixel's distribution $\mathcal{D}$ can be addressed using the framework developed in [57] in which the authors expose a method to incorporate physical realizability constraints at training time into additional regularizers.

C. RESULTS IN THE NOISY FRAMEWORK

We set the characteristics of our noisy photosensor to $(b, c, \sigma_{\mathrm{rd}}) := (14, 2 \cdot 10^4, 40)$ which are typical values for a medium-quality photosensor. We experimented with two values for the gain factor: $g = 1.0$ (low-level noise), and $g = 0.33$ (high-level noise). These are the values used in [16]. Like our experiments in the noiseless case, we trained and evaluated the architecture for a number of shots $N \in \{1, 2, 3\}$. To speed up training, we initialized the weights of the proximal operators to those learned in the noiseless case, and reset the learned values for μ_k to 1. This allowed us to halve the time needed for training. All hyper-parameters remain the same. We found it useful to apply on-the-fly random γ-correction on the light fields patches in the training set, by randomly drawing γ from $\mathcal{U}(0.6, 1.0)$.This allows us to avoid having training data with the same intensity levels as the validation and test sets. While discrepancy in average intensity does not seem to impede the ability of the network to generalize to brighter light fields in the noiseless case, it becomes a problem in the noisy case, most likely because the noise depends on the intensity levels. When using training data with low intensity levels, the trained model is not sufficiently exposed to high levels of noise.



(a) *Noiseless* **(b)** *Low-level noise* **(c)** *High-level noise* **(d)** *Ground truth*

FIGURE 10: **Reconstruction error of the central view of the light field** *Seahorse* **from the Kalantari test dataset [58], for different levels of noise.** The number of shots is $N = 3$.

Table 4 gives a quantitative comparison of the performance in the noisy case depending on the level of noise and the number of shots used for reconstruction. Table 4 also includes results in the single-shot case, using the same architecture, but without learning the CFA and CCM. It is worth noticing that learning the CFA and CCM along with the weights of the reconstruction network yields significant improvement, especially when the noise level is high. Table 5 shows examples of learned CCM and CFA for different levels of noise and different numbers of shots. We can see that the transmittance of the pixels of both the CFA and CCM tends to increase as the gain factor g decreases. Figure 10 shows the reconstruction error on the central view for the different levels of noise.

Noise	Low-level				High-level			
Number of shots	1*	1	2	3	1*	1	2	3
Buttercup	32.70	33.84	34.76	35.23	29.84	32.38	33.11	33.67
Cars	31.35	31.93	33.47	34.23	29.13	31.00	31.86	32.65
Orchids	32.56	33.27	34.89	35.48	30.13	31.96	32.94	33.76
Rock	32.51	32.86	35.48	36.24	29.71	31.30	32.85	34.09
Seahorse	33.94	34.18	36.22	36.90	31.57	32.82	34.47	35.32
Tulips	42.03	43.93	44.96	45.49	37.96	41.70	42.34	42.94
White rose	32.71	33.49	35.05	35.65	30.31	32.36	33.32	33.96
Average	33.97	34.78	36.40	37.03	31.24	33.36	34.41	35.26

TABLE 4: **PSNR comparison (dB) in the noisy framework for different levels of noise and numbers of acquisitions.** The first four columns correspond to a level of noise given by $g = 1.0$ and the last four columns to $g = 0.33$. The symbol * indicates a *fixed CCM and CFA*.

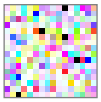
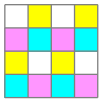
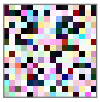
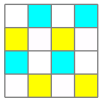
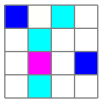
Noise	Low-level		High-level	
Component	CCM	CFA	CCM	CFA
1 shot				
2 shots				
3 shots				

TABLE 5: **Learned color filter arrays and color-coded masks**. 'Low-level' corresponds to $g = 1.0$ while 'High-level' corresponds to $g = 0.33$. The CCM displayed are random samples.

VII. LIMITATIONS

Whereas our approach performs well when the amount of noise is moderate, a significant reduction of the signal-to-noise ratio can greatly degrade the quality of the reconstruction. This makes our setup less efficient when the amount of light is very low. In addition, due to the convolutional nature of the network, light fields with large disparities (*i.e.* exceeding the size of the receptive field of the CNN) are usually not well recovered.

VIII. CONCLUSION

In this paper, we have presented a new deep architecture, based on unrolled optimization with learned priors, for the reconstruction of compressively acquired light fields *via* color-coded masks in the presence of color filter arrays. This architecture leverages the power of the algorithm unrolling paradigm, and works with an arbitrary number of shots. We have shown that this method improves the state-of-the-art for this acquisition framework by several dBs, comparing favorably to both traditional and deep methods [14]–[16]. We have presented an efficient closed-form data-term minimization layer that is shown to substantially improve the reconstruction quality, while allowing the joint learning of the coded mask and color filter array, along with the weights of the network. In addition, we have presented a new realistic acquisition model together with a method to compute its modulation field. Finally, we have shown that our approach was robust to realistic levels of noise, an important consideration in regard to practical applications.

REFERENCES

[1] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Linear volumetric focus for light field cameras," *ACM Trans. Graph.*, vol. 34, no. 2, mar 2015. [Online]. Available: https://doi.org/10.1145/2665074

[2] M. Levoy, B. Chen, V. Vaish, M. Horowitz, I. McDowall, and M. Bolas, "Synthetic aperture confocal imaging," *ACM Trans. Graph.*, vol. 23, no. 3, p. 825–834, aug 2004. [Online]. Available: https://doi.org/10.1145/1015706.1015806

[3] M. Levoy, B. Chen, V. Vaish, M. Horowitz, I. McDowall, and M. T. Bolas, "Synthetic aperture confocal imaging," in *SIGGRAPH 2004*, 2004.

[4] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, "Light field microscopy," in *ACM SIGGRAPH 2006 Papers*, ser. SIGGRAPH '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 924–934. [Online]. Available: https://doi.org/10.1145/1179352.1141976

[5] J. Shi, X. Jiang, and C. Guillemot, "A framework for learning depth from a flexible subset of dense and sparse light field views," *IEEE Transactions on Image Processing*, vol. 28, pp. 5867–5880, 2019.

[6] Y. Wang, L. Wang, G. Wu, J. Yang, W. An, J. Yu, and Y. Guo, "Disentangling light fields for super-resolution and disparity estimation," *IEEE Transactions on Pattern Analysis Machine Intelligence*, no. 01, pp. 1–1, feb 5555.

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

[8] E. Miandji, M. Emadi, J. Unger, and E. Afshari, "On probability of support recovery for orthogonal matching pursuit using mutual coherence," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1646–1650, Nov. 2017.

[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*, vol. 60, no. 9, pp. 5702–5715, Sep. 2014.

[10] J. Tropp, "Greed is good: algorithmic results for sparse approximation," *Information Theory, IEEE Transactions on*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[11] 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)*, Nov 2009, pp. 2337–2340.

[12] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive Light Field Photography using Overcomplete Dictionaries and Optimized Pro-

jections," *ACM Transactions on Graphics (Proceedings of SIGGRAPH)*, vol. 32, no. 4, pp. 46:1–46:12, 2013.
[13] 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)*, Jun 2018, pp. 226–230.
[14] H.-N. Nguyen, E. Miandji, and C. Guillemot, "Multi-mask camera model for compressed acquisition of light fields," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 191–208, 2021.
[15] O. Nabati, D. Mendlovic, and R. Giryes, "Fast and accurate reconstruction of compressed color light field," *2018 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–11, 2018.
[16] G. Le Guludec, E. Miandji, and C. Guillemot, "Deep light field acquisition using learned coded mask distributions for color filter array sensors," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 475–488, 2021.
[17] Y. 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*, vol. 1. IEEE, Nov. 1993, pp. 40–44.
[18] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *International Conference on Machine Learning (ICML)*, 2017, pp. 537–546.
[19] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *International Conf. on international conference on machine learning*, 2010, pp. 399–406.
[20] S. Diamond, V. Sitzmann, F. Heide, and G. Wetzstein, "Unrolled optimization with deep priors," *arXiv preprint arXiv:1705.08041*, 2017.
[21] M. Mardani, Q. Sun, S. Vasawanal, V. Papyan, H. Monajemi, J. Pauly, and D. Donoho, "Neural proximal gradient descent for compressive imaging," in *NeurIPS*, 2018.
[22] Y. Yang, J. Sun, H. Li, and Z. Xu, "Deep admm-net for compressive sensing mri," in *International Conf. on neural information processing systems*, 2016, pp. 10–18.
[23] J. Dong, S. Roth, and B. Schiele, "Deep wiener deconvolution: Wiener meets deep learning for image deblurring," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
[24] U. Schmidt and S. Roth, "Shrinkage fields for effective image restoration," in *IEEE Conf. on computer vision and pattern recognition*, 2014, pp. 2774–2781.
[25] 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)*, Oct 2020, pp. 278–294.
[26] G. Wetzstein, I. Ihrke, D. Lanman, and W. Heidrich, "State of the art in computational plenoptic imaging," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010.
[27] C. Liang, T. Lin, B.-Y. Wong, C. Liu, and H. Chen, "Programmable Aperture Photography: Multiplexed Light Field Acquisition," *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 55:1–55:10, Aug 2008.
[28] E. Miandji, J. Kronander, and J. Unger, "Compressive Image Reconstruction in Reduced Union of Subspaces," *Computer Graphics Forum*, vol. 34, no. 2, pp. 33–44, May 2015.
[29] S. D. Babacan, R. Ansorge, M. Luessi, P. R. Mataran, R. Molina, and A. K. Katsaggelos, "Compressive Light Field Sensing," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4746–4757, Dec 2012.
[30] 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.
[31] M. Guo, J. Hou, J. Jin, J. Chen, and L. Chau, "Deep spatial-angular regularization for light field imaging, denoising, and super-resolution," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 01, pp. 1–1, jun 5555.
[32] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, pp. 1207–1223, 2005.
[33] 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.
[34] 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)*, Jul 2017, pp. 1277–1286.
[35] J. H. R. Chang, C. Li, B. Póczos, and B. Kumar, "One network to solve them all — solving linear inverse problems using deep projection models," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5889–5898.
[36] H. Gupta, K. H. Jin, H. Q. Nguyen, M. T. McCann, and M. Unser, "Cnn-based projected gradient descent for consistent ct image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, p. 1440–1453, 2018.
[37] T. Meinhardt, M. Moeller, C. Hazirbas, and D. Cremers, "Learning proximal operators: Using denoising networks for regularizing inverse imaging problems," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1799–1808.
[38] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, p. 1256–1272, 2017.
[39] D. Gilton, G. Ongie, and R. Willett, "Deep equilibrium architectures for inverse problems in imaging," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 1123–1133, 2021.
[40] S. Bai, J. Z. Kolter, and V. Koltun, "Deep equilibrium models," in *NeurIPS*, 2019.
[41] D. Lanman, R. Raskar, A. Agrawal, and G. Taubin, "Shield fields: Modeling and capturing 3d occluders," *ACM Transactions on Graphics (TOG)*, vol. 27, no. 5, pp. 1–10, 2017.
[42] B. E. Bayer, "Color Imaging Array," U.S. Patent 3 971 065, Jul 20, 1976.
[43] L. Jiang, K.-J. Kim, F. M. Reininger, S. Jiguet, and S. Pau, "Microfabrication of a color filter array utilizing colored su-8 photoresists," *Appl. Opt.*, vol. 59, no. 22, pp. G137–G145, Aug 2020. [Online]. Available: http://www.osapublishing.org/ao/abstract.cfm?URI=ao-59-22-G137
[44] C. Williams, G. S. D. Gordon, T. D. Wilkinson, and S. E. Bohndiek, "Grayscale-to-color: Scalable fabrication of custom multispectral filter arrays," *ACS Photonics*, vol. 6, no. 12, pp. 3132–3141, 2019. [Online]. Available: https://doi.org/10.1021/acsphotonics.9b01196
[45] 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*, vol. 3, no. 3, pp. 384–397, 2017.
[46] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.
[47] W. Dong, P. Wang, W. Yin, G. Shi, F. Wu, and X. Lu, "Denoising prior driven deep neural network for image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 2305–2318, 2019.
[48] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep cnn denoiser prior for image restoration," in *IEEE Conf. on computer vision and pattern recognition*, 2017, pp. 3929–3938.
[49] HandWiki, "Woodbury matrix identity — handwiki,," 2020, [Online; accessed 15-June-2021]. [Online]. Available: https://handwiki.org/wiki/index.php?title=Woodbury_matrix_identity&oldid=17480
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
[51] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *4th International Conference on Learning Representations, ICLR 2016*, 2016. [Online]. Available: http://arxiv.org/abs/1511.07289
[52] R. Shah, G. Wetzstein, A. Sunder Raj, and M. Lowney. (2018) Stanford lytro light field archive. [Online]. Available: http://www.lightfields.stanford.edu/LF2016.html
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.
[54] M. Le Pendu, X. Jiang, and C. Guillemot, "Light field inpainting propagation via low rank matrix completion," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1981–1993, 2018.
[55] J. Shi, X. Jiang, and C. Guillemot, "A framework for learning depth from a flexible subset of dense and sparse light field views," *IEEE Transactions on Image Processing*, vol. PP, 06 2019.
[56] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," 03 2017, pp. 19–34.
[57] J. Bacca, T. Gelvez Barrera, and H. Arguello, "Deep coded aperture design: An end-to-end approach for computational imaging tasks," *IEEE Transactions on Computational Imaging*, vol. PP, pp. 1–1, 10 2021.
[58] 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)*, vol. 35, no. 6, 2016.

GUILLAUME LE GULUDEC received the M.S. degree in mathematics and fundamental computer science from Paris-Saclay University in 2016, and the Engineering degree from Télécom ParisTech in 2017. From 2017 to 2019 he has worked as a software developer, and has been a research engineer with Inria Rennes since 2020. His research interests include deep learning and methods for solving inverse problems applied to light field processing.

CHRISTINE GUILLEMOT , IEEE fellow, is Director of Research at INRIA. She holds a Ph.D. degree from ENST (Ecole Nationale Superieure des Telecommunications) Paris, and an Habilitation for Research Direction from the University of Rennes. From 1985 to Oct. 1997, she has been with FRANCE TELECOM, where she has been involved in various projects in the area of image and video coding and processing for TV, HDTV and multimedia. From Jan. 1990 to mid 1991, she has worked at Bellcore, NJ, USA, as a visiting scientist. Her research interests are signal and image processing, and computer vision. She has served as Associate Editor for IEEE Trans. on Image Processing (from 2000 to 2003, and from 2014-2016), for IEEE Trans. on Circuits and Systems for Video Technology (from 2004 to 2006), and for IEEE Trans. on Signal Processing (2007-2009). She has served as senior member of the editorial board of the IEEE journal on selected topics in signal processing (2013-2015) and is currently senior area editor of IEEE Trans. on Image Processing.