Multi-Mask Camera Model for Compressed Acquisition of Light Fields

Hoai Nam Nguyen, Ehsan Miandji, Christine Guillemot

Multi-Mask Camera Model for Compressed Acquisition of Light Fields

Hoai-Nam Nguyen, Ehsan Miandji, Christine Guillemot, *Fellow, IEEE*
Inria Rennes – Bretagne-Atlantique
263 Avenue Général Leclerc, 35042 Rennes Cedex, France

***Abstract*—We present an all-in-one camera model that encompasses the architectures of most existing compressive-sensing light-field cameras, equipped with a single lens and multiple amplitude coded masks that can be placed at different positions between the lens and the sensor. The proposed model, named the equivalent multi-mask camera (EMMC) model, enables the comparison between different camera designs, e.g using monochrome or CFA-based sensors, single or multiple acquisitions, or varying pixel sizes, via a simple adaptation of the sampling operator. In particular, in the case of a camera equipped with a CFA-based sensor and a coded mask, this model allows us to jointly perform color demosaicing and light field spatio-angular reconstruction. In the case of variable pixel size, it allows to perform spatial super-resolution in addition to angular reconstruction. While the EMMC model is generic and can be used with any reconstruction algorithm, we validate the proposed model with a dictionary-based reconstruction algorithm and a regularization-based reconstruction algorithm using a 4D Total-Variation-based regularizer for light field data. Experimental results with different reconstruction algorithms show that the proposed model can flexibly adapt to various sensing schemes. They also show the advantage of using an in-built CFA sensor with respect to monochrome sensors classically used in the literature.**

***Index Terms*—Light Field imaging, camera models, compressed sensing, regularization, inverse problems**

I. Introduction

Light field imaging is gaining in popularity for a number of applications, going from photo-realistic image-based rendering [1], [2], [3] to computational photography [4], [5], [6], [7], glass-free 3D displays [8], [9], and computer vision applications [10]. A light field is a sampled version of the 7D plenoptic function that describes the intensity of the light rays interacting with the scene as received by an observer at every point in space, along any direction of gaze, for all times and every wavelength [11]. In computer vision, light fields are commonly represented as four-dimensional (4D) functions with two spatial and two angular (or directional) dimensions [1], [2]. In this point-of-view, they can be seen as collections of two-dimensional (2D) images, which are called viewpoints or sub-aperture images or angular images, taken from different vantage points. The difference between viewpoints of the same scene enables not only reconstruction of 3D objects but also photo rendering with controlled depth-of-field.

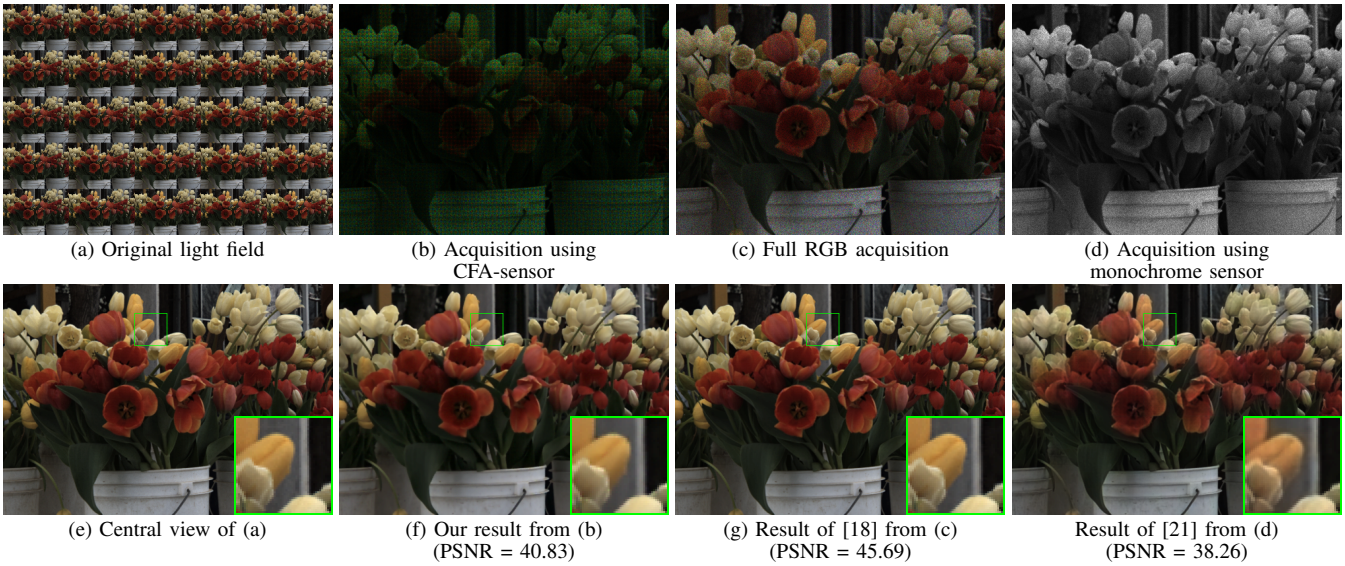
Over the past decades, a number of acquisition devices have been developed by both academic and industrial groups, with the aim of efficiently capturing 4D light fields. Existing devices can be classified into two categories: multi-view imagers which directly acquire different viewpoints using optical and mechanical setups, and multi-view coders which encode angular information of 4D light fields onto 2D sensor images and then "compute" original viewpoints. The first category includes imaging systems involving single cameras mounted on mechanical gantries moving over a spherical or planar surface [1], and multiple cameras arranged in an array fashion (also called camera arrays) [12], [13]. While moving camera gantries have major difficulties for capturing non-static scenes, specially those with fast moving objects and illumination changes, camera arrays suffer from low angular resolution due to physical constraints. In addition, both of these light field capturing systems are often quite bulky and not suitable for popular consumer uses.

In comparison with these cumbersome setups, compact devices have been designed by modifying the architecture of conventional cameras in order to capture the light field passing through the camera main lens. Early prototypes commonly known as "plenoptic cameras" [14], [15] use micro-lens arrays (MLA) placed on the optical path before the camera sensor to separate incoming light rays by their incident angle. Thus, this separation allows for acquiring multiple viewpoints of the scene. Alternatively, several recent light-field camera designs consider coded masks instead of an MLA to modulate 4D light fields into 2D projections that are captured by the camera sensor, and use a reconstruction algorithm to restore unmodulated light fields from their projections [16], [17], [18]. Note that most mask-based cameras are actually multi-view coders; in contrast, MLA-based (plenoptic) cameras belong to the class of multi-view imagers.

While an MLA-based camera can be modelled as a virtual or equivalent camera array (ECA) [19], there is no published work providing a similar equivalent system for mask-based prototypes. We introduce, in this work, a unifying model that represents any single-lens light-field camera with coded masks as an equivalent multi-mask camera (EMMC). The EMMC formalism enables the comparison of different camera designs and acquisition settings, e.g. using monochrome or CFA-based sensors, using single or multiple acquisitions, varying the pixel size. In particular, the architecture combining an attenuation mask and a CFA-equipped sensor can be seen as a dual-mask model: one is the coded mask which multiplexes angular (or directional) information and the other corresponds to the CFA pattern which samples color information. Preliminary results on joint demosaicing and light field reconstruction with CFA sensors have been published in [20]. Also, please note that the proposed camera model concerns only (amplitude) mask-based cameras and does not include camera designs using phase plates or micro-lens arrays.

Regarding light field reconstruction methods, the EMMC model, in which the sampling operator modeling the coded projection process of light fields can be flexibly computed according to the camera design changes, can be used with any reconstruction algorithm. While we first validate the model with existing dictionary-based reconstruction algorithms, we

This work was supported by the EU H2020 Research and Innovation Programme under grant agreement No 694122 (ERC advanced grant CLIM).



(a) Original light field (b) Acquisition using CFA-sensor (c) Full RGB acquisition (d) Acquisition using monochrome sensor

(e) Central view of (a) (f) Our result from (b) (PSNR = 40.83) (g) Result of [18] from (c) (PSNR = 45.69) Result of [21] from (d) (PSNR = 38.26)

Fig. 1: **Reconstruction of a color light field from a single 2D coded projection captured in various acquisition scenarios.** The original light field (a) is the "Tulip" scene (5×5 views) from the Stanford Lytro Illum dataset [22]. The coded projection is computed with a RGBW mask as in [21], and it is captured in three different ways: acquisition with CFA-sensor (b), full (RGB) color acquisition (c) and acquisition with monochrome sensor (d). Experimental results show that the reconstruction with our variational method from CFA-based acquisition is comparable to the result obtained with a state-of-the-art dictionary-based method [18] on full-color acquisition. In contrast, the combination of deep-learning-based algorithm [21] and monochrome acquisition produces a result with incorrect-color reconstruction.

also develop a generic variational algorithm by extending the total variation regularization for light field data. Experimental results on both synthetic and real light field data, show that our camera model can flexibly adapt to various acquisition schemes. They also show the advantage of using an in-built CFA sensor with respect to monochrome sensors classically used in the literature. In fact, when using the CFA sensor combined with a coded mask, the proposed model allows us to perform a joint demosaicing and light field reconstruction. Also, the model gives the flexibility of increasing the number of shots, thus giving better results compared to reconstructions from one-shot acquisitions.

In summary, the contributions of this paper are the following:

- The construction of an Equivalent Multi-Mask Camera (EMMC) model which unifies most existing single-lens mask-based light field cameras, and which allows flexible configuration of a variety of sensing schemes related to the number and the positions of coded masks, monochrome and color filter array (CFA)-based sensors, and the pixel size.
- The application of EMMC model in multi-shot light field reconstruction, joint light field reconstruction and demosaicing using a CFA-equipped sensor and coded masks, and joint light field reconstruction and super-resolution in the case of variable pixel size.
- The derivation of an image formation model using the proposed EMMC model to describe various acquisition schemes.
- The validation of the proposed model with various acquisition schemes using dictionary-based reconstruction methods as well as an approach combining a dictionary based reconstruction with a K-order differential regularization.

II. Related work

This section gives an overview of main computational light field camera designs, with a focus on camera designs aiming at capturing a whole light field with a single sensor, e.g. micro-lens-based and mask-based cameras. A taxonomy of computational light field cameras, according to the approach used to project the 4D light field onto a 2D sensor, can be found in [23].

MLA-based cameras

The first single-lens plenoptic camera has been introduced in [14], using an array of pinholes in front of the sensor. Following principles of integral imaging pioneered by Lippmann [24], each pinhole acts as a camera that creates a micro-image of the main lens aperture on a small area of the sensor. An array of micro-lenses can be considered instead of the pinhole array for better light efficiency. The authors of [14] also proposed an elaborated design which places a field lens in front of the MLA to reduce aberrations and place a relay lens right after the MLA to separate it from the sensor. A decade later, by modifying the original setup in [14], Ren Ng has developed the first hand-held light field camera [15], [4] that slightly modifies the original setup of [14]. A light transport framework for lenslet light field cameras is proposed in [25] taking into account non uniform angular pixel sensitivity to understand the limits of lenslet-based light field cameras. The authors in [26] also take into account lens aberrations in the light field camera design.

Mask-based cameras

Programmable aperture approaches have also been considered to sequentially capture subsets of light rays, hence time-multiplexing 2D slices of the 4D light field on the sensor, using a programmable non-refractive mask at the aperture [27], and exploiting the fast multiple-exposure feature of digital sensors. Shield fields are introduced in [28] to quantify the 4D attenuation of light rays due to occluders in mask-based camera designs. A good overview of the above camera designs can be found in [29].

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. The application of the compressive sensing framework to the problem of light field acquisition thus led to novel camera architectures, referred to as compressive coded aperture light field cameras. Thanks to the use of a coded mask, instead of recording a spatial multiplex of 2D slices of the light field, 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 [6], [7] 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 [30] 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 [17] propose a camera architecture that records optically coded projections on a single image sensor, while the authors in [31] and [18] use respectively a random binary mask or a moving colour coded mask affixed to the sensor 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 [17], [18] or an ensemble of 2D separable dictionaries [31]. Light field reconstruction from a sparse set of measurements is an inverse problem that can also be efficiently solved using deep learning techniques [32], [33], [21], [34]. The authors in [32], [33], [21] assume a pre-defined mask pattern and propose convolutional neural network architectures to reconstruct the light field from the set of measurements given the coded mask. In contrast, the authors in [34] 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. The higher resolutions achieved thanks to compressive acquisition is shown in [35] to be useful in light-field microscopy for 3D neural activity recording. Although not related to acquisition, we would like to mention that similar principles using masks or light-attenuating layers have also been considered to realize compressive light field displays [9].

Other single sensor camera designs

The authors in [36] also proposed a miniaturized and integrated camera array where multiple cameras share the same sensor, while in [37], the capture is done through a spray of water droplets on an acrylic surface placed in front of the camera [37]. Each drop of water produces a distorted view of the scene. A camera architecture is proposed in [38] allowing us to switch between a high resolution 2D image capture and a light field capture. This camera uses angular sensitive pixels (ASP) that allow for angular radiance information to be captured without the need for additional microlenses or light-blocking masks.

III. Imaging model: Background

In this section, we recall basic concepts of light field parameterization and imaging based on geometrical optics for conventional cameras featuring a main lens placed on one side and a photo-sensitive surface (also called "sensor") on the opposite side. While the plane at which the camera lens is located is called the "lens plane", the plane containing the photo-sensitive surface is often referred to as the "sensor plane". We then establish the link between the light field induced by a world 3D scene and the image of this scene through the main lens as a projection onto the sensor plane.

Symbols	Description
$\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{x}, \boldsymbol{u} \in \mathbb{R}^2$	Points on 2D planes
$L_{\boldsymbol{Z}}(\boldsymbol{X}, \boldsymbol{U}, \lambda)$	Radiance of the light rays passing through two points $\boldsymbol{X}$ and $\boldsymbol{U}$ located in two planes of depths $(Z, W) = \boldsymbol{Z}$ in the wavelength λ
$L_{\boldsymbol{z}_{\mathrm{cam}}}$	In-camera light field parameterized by the sensor and the lens planes with $\boldsymbol{z}_{\mathrm{cam}} = (z_{\mathrm{sensor}}, z_{\mathrm{lens}})$
$M(\boldsymbol{\xi}, \lambda)$	Value of the coded mask M at the position $\boldsymbol{\xi}$ in the wavelength λ
$\widetilde{I}_{\lambda}$	Angular-compressed projection of the light field in the wavelength λ
$\overline{I}$	Spectro-angular-compressed projection of the light field (all color channels are merged)
$\boldsymbol{k} = (k_1, k_2) \in \mathbb{Z}^2$	Discrete coordinates on the sensor.
$\boldsymbol{I}_{\boldsymbol{k}}^{(s)}$	Intensity of the pixel of width $\triangle_p$ and center $\boldsymbol{c}_{\boldsymbol{k}}$ with the discrete cordinates $\boldsymbol{k}$ on the sensor, captured by the s^{th} acquisition
$\boldsymbol{\Phi} \in \mathbb{R}^{n_{\mathrm{s}} n \times 3\nu n}$	Coded projection matrix of $\nu = \nu_u \nu_v$ angular views, $n = n_x n_y$ being the number of spatial samples and n_{s} being the number of acquisitions.
$\mathbf{H} \in \mathbb{R}^{n_{\mathrm{s}} r n \times n_{\mathrm{s}} n}$	Sampling matrix modeling the sampling operations at the sensor level, $r = (\frac{\triangle_x}{\triangle_p})^2$ being the squared ratio between the spatial sampling step and the pixel width

TABLE I: **Notations and Symbols.**

A. Light field parameterization

We adopt the two-plane parameterization, in which any light field can be represented by a collection of light rays passing through two points on a pair of parallel planes. Given two constants Z and W, we denote by $L_{\boldsymbol{Z}}(\boldsymbol{X}, \boldsymbol{U}, \lambda)$ the radiance in the wavelength λ along a ray passing through the point $\boldsymbol{X} = (X, Y)$ on a plane of depth Z and the point $\boldsymbol{U} = (U, V)$ on a plane of depth W, where $\boldsymbol{Z} = (Z, W)$ is simply the depth vector of the two parameterization planes (see Fig.2). The couple (X, Y) (resp. (U, V)) represents the spatial coordinates (resp. angular coordinates) of the light field $L_{\boldsymbol{Z}}$. Here, the first reference plane (i.e. the $\boldsymbol{X}$-plane) is called the "focal plane" and the second one (i.e. the $\boldsymbol{U}$-plane) is called the "aperture plane" (which is often assumed to coincide with the "lens plane"). Common conventions often assume that the region between these two planes is a free space (i.e. there is no occlusion) to guarantee the preservation of radiance along light rays traveling in this region. Given a scene and a fixed camera position, the parameterization of the light fields can be extended for any two parallel planes of arbitrary depths (see Fig. 2 and [25], [39] for details). Hence different light fields representing the same (static) scene can be obtained by changing the depth position of the focal plane or the aperture plane, or even both of them.

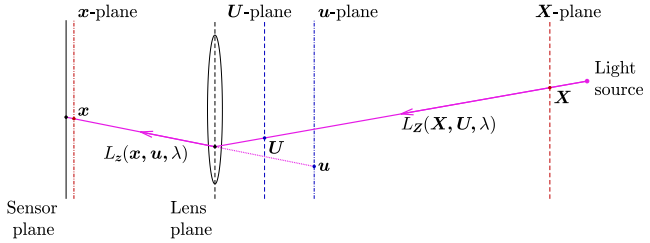


Fig. 2: **Lens imaging and light field parameterization.** The light field induced by the scene can be parameterized by any two parallel planes, e.g. the pair of X- and U-plane (black and blue dashed lines respectively). The camera lens transforms the "world light field" $L_{\boldsymbol{Z}}$ (also called the "outside light field") into the "in-camera light field" $L_{\boldsymbol{z}}$ which is a distorted version of $L_{\boldsymbol{Z}}$ with $L_{\boldsymbol{Z}}(\boldsymbol{X},\boldsymbol{U},\lambda) = L_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{u},\lambda)$. The $\boldsymbol{x}$-plane (resp. $\boldsymbol{u}$-plane) is the conjugate plane of the $\boldsymbol{X}$-plane (resp. U-plane) through the camera lens. Note that the $\boldsymbol{x}$-plane is not necessary coincide with the sensor plane of the camera.

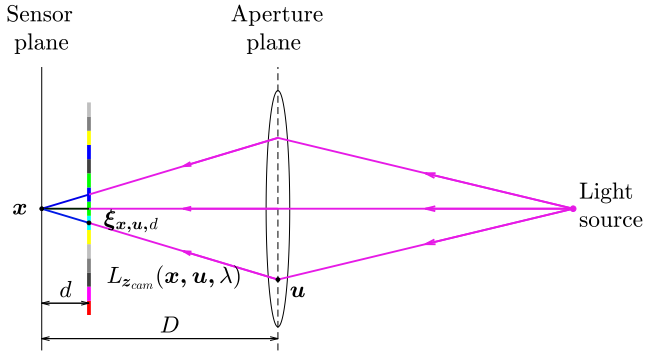


Fig. 3: **Standard mask-based light field camera model.** A coded mask is placed between the sensor and the main lens. Incoming light rays are filtered by the mask before they reach the sensor.

B. Image formation in a conventional camera

In this paper, we consider single-lens architectures and without loss of generality, we assume that the main lens of the camera is centered at origin (i.e. at the point $\mathbf{O} = (0,0,0)$). Let us consider a light ray starting from a point on an object of the scene and passing through two points $\boldsymbol{X}$ and $\boldsymbol{U}$ located on two planes of depth Z and W respectively. Due to refraction, the light ray bends after traveling through the camera lens and intersects the conjugate plane of the two reference planes at positions $\boldsymbol{x}$ and $\boldsymbol{u}$ as follows:

$$\boldsymbol{x} = \frac{f}{Z+f}\,\boldsymbol{X}\ , \qquad z = \frac{fZ}{Z+f}\ ,$$
$$\boldsymbol{u} = \frac{f}{W+f}\,\boldsymbol{U}\ , \qquad w = \frac{fW}{W+f}\ ,$$

where z (resp. w) denotes the depth of the conjugate plane of the $\boldsymbol{X}$-plane (resp. $\boldsymbol{U}$-plane), and where f denotes the camera focal length. Or in other words, the camera lens transforms the "world light field" $L_{\boldsymbol{Z}}$ (also called the "outside light field") into the "in-camera light field" $L_{\boldsymbol{z}}$ which is a distorted version of $L_{\boldsymbol{Z}}$ with $L_{\boldsymbol{Z}}(\boldsymbol{X},\boldsymbol{U},\lambda) = L_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{u},\lambda)$ (see Fig.2). When a translucent screen is placed exactly on the $\boldsymbol{x}$-plane (i.e. the sensor plane coincides with the conjugate plane of the $\boldsymbol{X}$-plane), classical radiometry states that the irradiance in the wavelength λ at a position $\boldsymbol{x}$ on the screen is equal to the following integral [4]:

$$I_\lambda(\boldsymbol{x}) = \int_{\Theta} L_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{u},\lambda)\,d\boldsymbol{u}$$
$$= \frac{f}{W+f}\int_{\frac{W+f}{f}\Theta} L_{\boldsymbol{Z}}\Big(\frac{Z+f}{f}\boldsymbol{x},\boldsymbol{U},\lambda\Big)d\boldsymbol{U}\ , \qquad (1)$$

where $\Theta \subset \mathbb{R}^2$ is the camera aperture domain. This equation represents the photograph formation in a conventional single-lens camera. It is nothing else than the projection of the 4D light field onto a 2D image by summing all rays intersecting the sensor plane at the same position.

IV. Revisiting acquisition using coded masks

In this section, we study the image formation on a slightly modified conventional camera architecture, in which one or multiple (coded) masks can be inserted between the main lens and the sensor (see Fig.3). In such an architecture, we denote by $L_{\boldsymbol{z}_{\text{cam}}}$ the in-camera light field parameterized by the sensor plane and the lens plane (here $\boldsymbol{z}_{\text{cam}} = (z_{\text{sensor}}, z_{\text{lens}})$, where z_{sensor} and z_{lens} is the depth of the sensor plane and the main lens plane respectively). Also, let $D = |z_{\text{sensor}} - z_{\text{lens}}|$ be the distance between two parameterization planes. We assume furthermore that $L_{\boldsymbol{z}_{\text{cam}}}$ is simply zero beyond the physical bounds of the camera sensor (denoted by $\Omega \in \mathbb{R}^2$) and the aperture Θ.

If the sensor is a translucent screen and a coded mask $M(\boldsymbol{\xi},\lambda)$ (with feature size $\triangle_M$, i.e. the size of each element or pixel of the mask) is placed at a distance d from the sensor plane ($0 \le d \le D$) in a way that $\frac{d}{\triangle_M}$ is sufficiently small with respect to the visible spectrum (corresponding to 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 [40]. Then, the imaging equation can be rewritten as [17], [18]:

$$\widetilde{I}_\lambda(\boldsymbol{x}) = \int_{\Theta} M(\boldsymbol{\xi}_{\boldsymbol{x},\boldsymbol{u},d},\lambda)\ L_{\boldsymbol{z}_{\text{cam}}}(\boldsymbol{x},\boldsymbol{u},\lambda)\ d\boldsymbol{u}\ , \qquad (2)$$

where $\boldsymbol{\xi}_{\boldsymbol{x},\boldsymbol{u},d} = \boldsymbol{u} + \frac{D-d}{D}(\boldsymbol{x}-\boldsymbol{u})$. Each image $\widetilde{I}_\lambda$ is called "coded projection" of the original light field $L_{\boldsymbol{z}_{\text{cam}}}$ in the color channel λ and can be seen as an angular-compressed version of the set of angular images $\{L_{\boldsymbol{z}_{\text{cam}}}(\cdot,\boldsymbol{u},\lambda)\}_{\boldsymbol{u}\in\Theta}$. Note that Eqn. (1) is a particular case of Eqn. (2) when $M = 1$, i.e. when the mask is fully transparent.

Let us replace the translucent screen in the camera architecture by a simple monochrome sensor array that receives all incoming lights but does not distinguish the light colors. Accordingly, the intensity of an idealized infinitesimal pixel (i.e. the pixel size is considered as a point particle) located at the position $\boldsymbol{x}$ on the sensor plane can be computed as follows:

$$\overline{I}(\boldsymbol{x}) = \int_{\Lambda} \widetilde{I}_\lambda(\boldsymbol{x})\,d\lambda = \int_{\Theta\times\Lambda} M(\boldsymbol{\xi}_{\boldsymbol{x},\boldsymbol{u},d},\lambda)\,L_{\boldsymbol{z}_{\text{cam}}}(\boldsymbol{x},\boldsymbol{u},\lambda)\,d\boldsymbol{u}\,d\lambda\ , \qquad (3)$$

where Λ is the range of light wavelengths that can be captured by the sensor. One can remark that if the mask M is not mounted directly on the sensor (i.e. $d \neq 0$) as suggested in [21], all color channels of $\{\widetilde{I}_\lambda\}_{\lambda\in\Lambda}$ are merged to form the "gray-scale" image $\overline{I}$ which contains only luminance information. In this case, both directional (angular) and spectral (color) information are compressed. Consequently, reconstructing the light field $L_{\boldsymbol{z}_{\text{cam}}}$ from its spectro-angular-compressed projection $\overline{I}$ is more challenging since it requires not only the estimation of the original viewpoints, but also the true color of each light field ray.

Fortunately, there are several acquisition schemes which permit avoiding the loss of all color information.

a) Multi-spectral sensing: It is possible to place color filters on the camera optical path for capturing coded pro-

jections in desired wavelengths. Each color channel can be obtained separately by considering the corresponding color filter. This enables multi-spectral imaging for arbitrary spectral (color) bands by the means of multiple (sequential) acquisitions; however, the total acquisition time increases proportionally to the number of colors that we want to capture. Another way to perform multi-spectral imaging without increasing the acquisition time is to use a prism to split incoming light beam into different color channels and use a large sensor array (or several small sensor arrays) to record each channel. Nevertheless, it requires major modifications of the camera architecture both in terms of optical setup and sensor implementation which are not convenient for popular consumers.

b) Monochromatic multi-shot acquisition: As in [18], one can also consider multiple shots, each of which is taken using a different random colored pattern. The color-coded mask is indeed changed to create a new random pattern for increasing the incoherence. In order to retrieve color information, each pattern is generated in the same (color) tint. More interestingly, multi-spectral sensing can be seen as a particular case of monochromatic multi-shot acquisition, in which the random pattern used for each shot is actually a color component of the given color-coded mask.

c) Using built-in CFA: An alternative solution for color light field imaging is to use sensors with a built-in color filter array (CFA) in order to sample the color information. Accordingly, captured images are color mosaics, in which the color of each pixel is defined by the color filter located at the pixel position. Note that, in an equivalent setting with monochrome sensors, it amounts to placing the corresponding CFA pattern right in front of the (monochrome) sensor array.

V. Equivalent Multi-Mask Camera model

In this section, we extend the conventional image formation model for single-lens single-mask cameras to more general situations, e.g. acquisition using multiple coded masks, multiple shots, and varying pixel sizes. We then introduce the notion of "Equivalent Multi-Mask Camera" that unifies various camera architectures and acquisition configurations. Using this EMMC formalism, we derive the imaging (sensing) matrix which maps a light field to a set of projected sensor images corresponding to a given situation. Also, we show the link between the proposed image formation model and the acquisition process on existing single-lens mask-based light field cameras.

A. Extended image formation model

The acquisition using a coded mask and a color sensor can be abstracted by an acquisition model using two coded masks and a monochrome sensor. In fact, the idea of using more than one mask has been introduced in [6], [7] with a dual-mask design: one mask is placed before the sensor and the other one at the aperture.

1) Multiple masks at varying positions: By extending this concept, we propose a camera model equipped with a main lens, a monochrome sensor and a set of n_{m} masks $\mathcal{M} = \{M_1, M_2, \ldots, M_{n_{\mathrm{m}}}\}$, in which each mask M_l is located between the camera sensor and aperture, at an arbitrary distance d_l from the sensor plane. Accordingly, the coded projection of the in-camera light field $L_{\boldsymbol{z}_{\mathrm{cam}}}$, through the masks $\mathcal{M}$ and captured by the sensor, reads:

$$I(\boldsymbol{x}) = \int_{\Theta\times\Lambda} \prod_{l=1}^{n_{\mathrm{m}}} M_l(\boldsymbol{\xi}_{\boldsymbol{x},\boldsymbol{u},d_l}, \lambda)\; L_{\boldsymbol{z}_{\mathrm{cam}}}(\boldsymbol{x}, \boldsymbol{u}, \lambda)\, d\boldsymbol{u}\, d\lambda \ . \quad (4)$$

The proposed multi-mask formalism does not only generalize the light field camera prototype introduced in [21], but it also represents an equivalent model for most existing mask-based light field cameras. The readers are referred to section V-C for a detailed comparison of different mask-based designs and their equivalent multi-mask camera.

2) With varying pixel supports and multiple acquisitions: In reality, although (digital) sensor pixels are relatively small comparing to the main lens or the sensor itself, they have a certain size that can not be reduced to an idealized point particle. Each pixel therefore measures the quantity of incident light over a bounded region called "pixel support" to form the pixel value. Note that the pixel support may have various shapes depending on the sensor design. For the sake of simplicity, we only consider, in this work, regular rectangular sensors with square pixels. Under this assumption, let $\triangle_p$ be the pixel size and let $\boldsymbol{k} = (k_1, k_2) \in \mathbb{Z}^2$ be the discrete coordinates of a sensor pixel. The support of this pixel is defined as the following rectangular domain:

$$\mathcal{C}_{\boldsymbol{k}} = \left[x_{k_1} - \frac{\triangle_p}{2}, x_{k_1} + \frac{\triangle_p}{2}\right[\times \left[y_{k_2} - \frac{\triangle_p}{2}, y_{k_2} + \frac{\triangle_p}{2}\right[\ ,$$

where $(x_{k_1}, y_{k_2}) = \boldsymbol{c}_{\boldsymbol{k}} \in \Omega$ denotes the pixel center. One can further envisage multiple acquisitions (shots) of the same scene and with the same camera position on the same sensor using different configurations of coded masks (e.g. changing mask position and pattern, inserting more masks or even removing some of them). Let n_{s} be the number of desired acquisitions, then the s-th captured image can be expressed as:

$$\boldsymbol{I}_{\boldsymbol{k}}^{(s)} = \int_{\mathcal{C}_{\boldsymbol{k}}\times\Theta\times\Lambda} \prod_{l=1}^{n_{\mathrm{m}}^{(s)}} M_l^{(s)}(\boldsymbol{\xi}_{\boldsymbol{x},\boldsymbol{u},d_l}^{(s)}, \lambda)\; L_{\boldsymbol{z}_{\mathrm{cam}}}(\boldsymbol{x}, \boldsymbol{u}, \lambda)\, d\boldsymbol{x}\, d\boldsymbol{u}\, d\lambda \ , \quad (5)$$

where $1 \leq s \leq n_{\mathrm{s}}$ and $\mathcal{M}^{(s)} = \{M_1^{(s)}, M_2^{(s)}, \ldots, M_{n_{\mathrm{m}}^{(s)}}^{(s)}\}$ is the set of $n_{\mathrm{m}}^{(s)}$ masks used for this acquisition. As each image $\boldsymbol{I}^{(s)}$ can be seen as a finite collection of sensor pixel values, the imaging equation Eqn. (5) establishes the link between the continuous light field $L_{\boldsymbol{z}_{\mathrm{cam}}}$ and the discrete form of its coded projections measured by the sensor. In practice, light field reconstruction algorithms tend to produce (from the set of measurements $\{\boldsymbol{I}_{\boldsymbol{k}}^{(s)}\}_{s,\boldsymbol{k}}$) an estimated discrete version of $L_{\boldsymbol{z}_{\mathrm{cam}}}$ (up to a given spatio-angular resolution) instead of its continuous counterpart. For this purpose, we introduce in section V-B the imaging (sensing) matrix representing the transformation of a discrete 4D color light field into captured sensor images of its coded projection by taking into account the spatial, angular, and spectral resolutions, as well as the size of sensor pixels and the number of acquisitions to be performed.

B. Discretized sensing matrix

Given two strictly positive constant $\triangle_x$ and $\triangle_u$, we assume that the in-camera light field $L_{\boldsymbol{z}_{\mathrm{cam}}}$ is sampled in red, green and blue colors (which correspond to the wavelength λ_{R}, λ_{G} and λ_{B} respectively) according to the grid of spatial and angular coordinates:

$$\mathcal{X} = \{\boldsymbol{x_i} = \boldsymbol{x_0} + \triangle_x \boldsymbol{i} \in \Omega \text{ for } \boldsymbol{i} = (i_1, i_2) \in \mathbb{N}^2\} \ ,$$
$$\mathcal{U} = \{\boldsymbol{u_j} = \boldsymbol{u_0} + \triangle_u \boldsymbol{j} \in \Theta \text{ for } \boldsymbol{j} = (j_1, j_2) \in \mathbb{N}^2\} \ ,$$

where $\boldsymbol{x_0} = (x_0, y_0)$ and $\boldsymbol{u_0} = (u_0, v_0)$ are some reference coordinates. Let $n = n_x n_y$ be the number of spatial samples and $\nu = \nu_u \nu_v$ the number of angular samples (also known

$$\underbrace{\begin{bmatrix}\boldsymbol{I}^{(1)}\\ \boldsymbol{I}^{(2)}\\ \vdots\\ \boldsymbol{I}^{(n_\mathrm{s})}\end{bmatrix}}_{\mathbf{I}} = \underbrace{\begin{bmatrix}\boldsymbol{H}^{(1)} & 0 & \dots & 0\\ 0 & \boldsymbol{H}^{(2)} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \boldsymbol{H}^{(n_\mathrm{s})}\end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix}\boldsymbol{\phi}^{(1,\mathrm{R},1)} & \dots & \boldsymbol{\phi}^{(\nu,\mathrm{R},1)} & \boldsymbol{\phi}^{(1,\mathrm{G},1)} & \dots & \boldsymbol{\phi}^{(\nu,\mathrm{G},1)} & \boldsymbol{\phi}^{(1,\mathrm{B},1)} & \dots & \boldsymbol{\phi}^{(\nu,\mathrm{B},1)}\\ \boldsymbol{\phi}^{(1,\mathrm{R},2)} & \dots & \boldsymbol{\phi}^{(\nu,\mathrm{R},2)} & \boldsymbol{\phi}^{(1,\mathrm{G},2)} & \dots & \boldsymbol{\phi}^{(\nu,\mathrm{G},2)} & \boldsymbol{\phi}^{(1,\mathrm{B},2)} & \dots & \boldsymbol{\phi}^{(\nu,\mathrm{B},2)}\\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{\phi}^{(1,\mathrm{R},n_\mathrm{s})} & \dots & \boldsymbol{\phi}^{(\nu,\mathrm{R},n_\mathrm{s})} & \boldsymbol{\phi}^{(1,\mathrm{G},n_\mathrm{s})} & \dots & \boldsymbol{\phi}^{(\nu,\mathrm{G},n_\mathrm{s})} & \boldsymbol{\phi}^{(1,\mathrm{B},n_\mathrm{s})} & \dots & \boldsymbol{\phi}^{(\nu,\mathrm{B},n_\mathrm{s})}\end{bmatrix}}_{\mathbf{\Phi}} \underbrace{\begin{bmatrix}\boldsymbol{L}^{(1,\mathrm{R})}\\ \vdots\\ \boldsymbol{L}^{(\nu,\mathrm{R})}\\ \boldsymbol{L}^{(1,\mathrm{G})}\\ \vdots\\ \boldsymbol{L}^{(\nu,\mathrm{G})}\\ \boldsymbol{L}^{(1,\mathrm{B})}\\ \vdots\\ \boldsymbol{L}^{(\nu,\mathrm{B})}\end{bmatrix}}_{\mathbf{L}}, \quad (6)$$

as "number of views"). We then denote by $\mathbf{L} \in \mathbb{R}^{3\nu n}$ the column vector which represents the sampled version of $L_{\boldsymbol{z}_\mathrm{cam}}$ according to the sampling grid $\mathcal{X} \times \mathcal{U}$ defined as:

$$\mathbf{L}^{\mathrm{c}}_{\boldsymbol{i},\boldsymbol{j}} = L_{\boldsymbol{z}_\mathrm{cam}}(\boldsymbol{x}_{\boldsymbol{i}}, \boldsymbol{u}_{\boldsymbol{j}}, \lambda_\mathrm{c}) \ ,$$

for every $\mathrm{c} \in \mathcal{C} = \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$. By construction, the size of the discrete light field $\mathbf{L}$ is $n_x \times n_y \times \nu_u \times \nu_v \times 3$. We denote the set of discrete spatial and angular coordinates by $\mathcal{I}$ and $\mathcal{J}$ respectively as follows:

$$\boldsymbol{i} \in \mathcal{I} = \{1, 2, \dots, n_x\} \times \{1, 2, \dots, n_y\} \subset \mathbb{N}^2 \ ,$$
$$\boldsymbol{j} \in \mathcal{J} = \{1, 2, \dots, \nu_u\} \times \{1, 2, \dots, \nu_v\} \subset \mathbb{N}^2 \ .$$

To simplify our notation, we adopt the following arrangement $\mathbf{L} = \{\boldsymbol{L}^{(j,\mathrm{c})}\}_{1 \leq j \leq \nu, \mathrm{c} \in \mathcal{C}}$, where $j = j_1 + \nu_u(j_2 - 1)$ and each $\boldsymbol{L}^{(j,\mathrm{c})} \in \mathbb{R}^n$ denotes the vectorized version of the j-viewpoint in the wavelength λ_c. Using this arrangement, the discrete form of the multi-mask-based multi-shot acquisition scheme in Eqn. (5) can be expressed as a matrix-vector multiplication, as shown in Eqn. (6). Here, $\mathbf{I} \in \mathbb{R}^{n_\mathrm{s} r n}$ denotes the collection of n_s acquired images, each $\boldsymbol{H}^{(s)} \in \mathbb{R}^{rn \times n}$ is the matrix representing the integration (summation) of incident light rays on the same sensor pixels for the s-th acquisition, $r = (\frac{\triangle_x}{\triangle_p})^2$ is the squared ratio between the spatial sampling step and the pixel width, each $\boldsymbol{\phi}^{(j,\mathrm{c},s)} = \prod_{l=1}^{n_\mathrm{m}^{(s)}} \boldsymbol{M}_l^{(j,\boldsymbol{c},s)} \in \mathbb{R}^{n \times n}$ is a sparse matrix and each $\boldsymbol{M}_l^{(j,\mathrm{c},s)} \in \mathbb{R}^{n \times n}$ is a diagonal matrix containing the coefficients in the wavelength λ_c of the mask $M_l^{(s)}$ on its diagonal. Moreover, $\mathbf{\Phi} \in \mathbb{R}^{n_\mathrm{s} n \times 3\nu n}$ is called the "coded projection matrix" containing the coefficients of implemented coded masks $\{\mathcal{M}^{(s)}\}_{1 \leq s \leq n_\mathrm{s}}$, and $\mathbf{H} \in \mathbb{R}^{n_\mathrm{s} r n \times n_\mathrm{s} n}$ is called the "sampling matrix" which models the sampling operations at the sensor level. By construction, the transform that relates the original (discrete) light field $\mathbf{L}$ to the sensor images $\mathbf{I}$ is given by:

$$\mathbf{\Psi} \ := \ \mathbf{H}\mathbf{\Phi} \ \in \ \mathbb{R}^{n_\mathrm{s} r n \times 3\nu n} \ . \quad (7)$$

One can easily see that the imaging matrix $\mathbf{\Psi}$ is sparse, due to the structure of $\mathbf{H}$ and $\mathbf{\Phi}$. Note that when $r = 1$, $\mathbf{H}$ becomes an identity matrix, and thus implying $\mathbf{\Psi} = \mathbf{\Phi}$, which represents the conventional compressive sensing (CS) framework of light fields as described in [16], [17], [18], [21]. Most early studies have proposed to reconstruct light fields with the same spatial resolution as of the captured sensor images. However, there is no published work that addresses the reconstruction of higher-spatial-resolution light fields from low-spatial-resolution coded projections. Note that the spatial sub-sampling technique used in [18] assumes that the spatial resolution of the light field coincides with the sensor resolution.

In practice, for real camera systems using amplitude-coded masks [30], [17], the imaging matrix of the whole system is estimated using the so-called "whiteboard" scenes. Details of such estimation procedure can be found in the supplement material of [17]. The estimation of the imaging matrix $\mathbf{\Psi}$ is also called as (PSF) "calibration".

C. Link with existing single-lens mask-based cameras

Most of existing single-lens camera designs represent special cases of Eqn. (5) by changing the number of masks n_m, the mask pattern M_l and the distance d_l between the mask M_l and the sensor plane. One trivial example is conventional (photography) cameras (which do not possess any coded mask) corresponding to $n_\mathrm{m} = 1$ and $M_1 = 1$. Regarding mask patterns, various classes of coded masks in the literature (such as broadband, sum-of-sines, random, monochrome, color etc.) can be easily modeled by using appropriate M_l functions.

a) Coded aperture cameras (Fig.4-a): Coded aperture cameras (also known as pupil plane coding cameras) use one mask located on the aperture plane of a traditional lens, i.e. corresponding to $n_\mathrm{m} = 1$ and $d_1 = D$. The mask pattern is either a sum of cosine signals in spatial domain (corresponding to a sum of Dirac delta functions in the Fourier domain) [5], a random pattern [16], or a learned pattern [34].

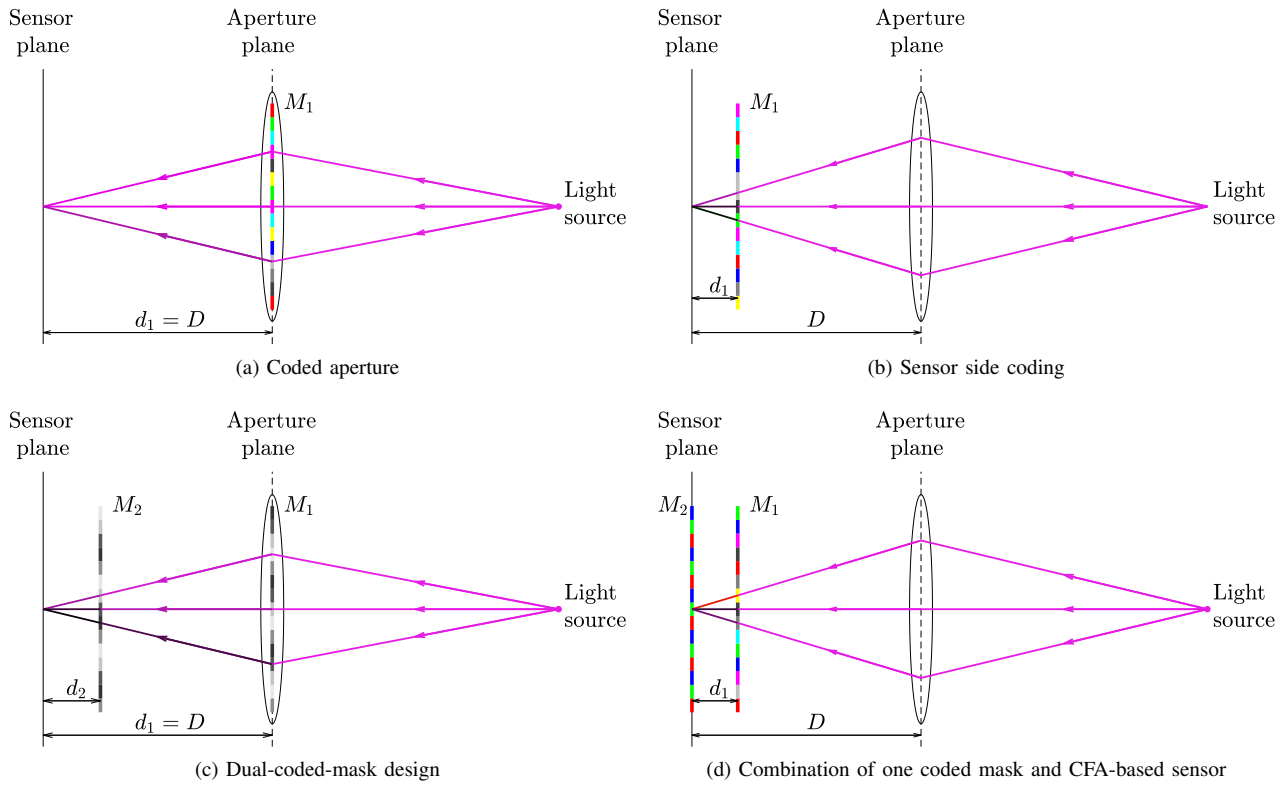
b) Sensor-side mask-based cameras (Fig. 4-b): These camera architectures (also known as sensor side coding cameras) use a mask placed on the sensor plane or in front of the sensor [17], [18], [21], hence usually setting $n_\mathrm{m} = 1$ and assuming $d_1 \ll D$. The mask patterns are in general random monochrome masks with Gaussian weights [17] or random color coded masks [18], [21]. When $d_1 = 0$, the mask plane coincides with the sensor plane. This setting allows modeling specific implementations at the sensor level, e.g. color filter arrays (such as the Bayer filter array) can be considered as particular color coded masks to encode color information into photo-sensitive sensors.

c) Dual-mask cameras: Previously, the two-mask camera proposed in [6], [7] is special case of the generic model of Eqn. (4), with $n_\mathrm{m} = 2$, $d_1 = D$ and $0 < d_2 < D$. In this camera model, one mask is placed in the aperture plane while the second mask is positioned in the optical path of the camera (see Fig. 4c). Recently, the authors of [20] consider a camera architecture combining a coded (attenuation) mask and a CFA-based sensor which is often implemented in standard consumer cameras. This architecture can be abstracted by an equivalent two-mask model: one mask is placed before the sensor ($d_1 > 0$) and the other one is mounted directly on the monochrome sensor (see Fig. 4d).

VI. Reconstruction algorithms

A. Dictionary-based reconstruction

The compressive sensing theory [42] relies on the assumption that the signal is sparse (or *compressible*) in some transform



(a) Coded aperture (b) Sensor side coding

(c) Dual-coded-mask design (d) Combination of one coded mask and CFA-based sensor

Fig. 4: **Mask-based light field camera architectures from the literature.** Coded aperture [41], [16]: (a) with $n_{\mathrm{m}} = 1$ and $d_1 = D$. Sensor side coding [17], [18]: (b) with $n_{\mathrm{m}} = 1$ and $d_1 \ll D$. Designs with two masks ($n_{\mathrm{m}} = 2$): (c) [6], [7] model using sum-of-sinusoid masks, and (d) combination of one coded mask and CFA-based sensor [20].

domain like wavelets, discrete cosine transform (DCT), or dictionaries learned from large datasets. In the context of compressed acquisition of light fields using coded masks, original light fields can be restored by solving a basis pursuit denoising (BPDN) problem [43], given an overcomplete dictionary as described in [17], [18]. We adapt this BPDN-based approach for the multi-mask-based multi-shot acquisition (Eqn. (6)).

Let us denote by $\mathbf{D} = \begin{bmatrix}\mathbf{D}^{(1)} & \mathbf{D}^{(2)} & \dots & \mathbf{D}^{(n_{\mathrm{d}})}\end{bmatrix} \in \mathbb{R}^{3\nu q \times n_{\mathrm{d}}}$ a light field dictionary, where $n_{\mathrm{d}} \geq 3\nu q$ is the number of dictionary atoms and each $\mathbf{D}^{(d)} = \{\boldsymbol{D}^{(j,\mathrm{c},d)}\}_{1\leq j\leq \nu, \mathrm{c}\in\mathcal{C}} \in \mathbb{R}^{3\nu q}$ is a (color 4D) light field atom of size $q_x \times q_y \times \nu_u \times \nu_v \times 3$ (with $q_x q_y = q$). Here, the first two dimensions are spatial dimensions, the third and fourth dimensions are angular dimensions and the last one is for the number of color channels. In practice, due to the large size of light field data, one usually trains dictionaries using small patches obtained by dividing full-size light fields over the spatial domain while including all angular and color dimensions to reduce the learning time of dictionaries. Using spatially small patches instead of whole light fields also allows increasing the number of training examples (or training samples) for the learning stage. Given a pre-trained dictionary, the reconstruction is first performed on light field patches of the same size as the dictionary atoms, and afterward reconstructed patches are aggregated to compute the light field by averaging pixels on overlapped regions.

For the light field $\mathbf{L} \in \mathbb{R}^{3\nu n}$, we consider a set of patches $\boldsymbol{\mathcal{L}} = \{\mathbf{L}^{(p)}\}_{1\leq p\leq n_{\mathrm{p}}}$ extracted from $\mathbf{L}$, where n_{p} denotes the number of patches and each $\mathbf{L}^{(p)} = \{\boldsymbol{L}^{(j,\mathrm{c},p)}\}_{1\leq j\leq \nu, \mathrm{c}\in\mathcal{C}} \in \mathbb{R}^{3\nu q}$ is a patch of size $q_x \times q_y \times \nu_u \times \nu_v \times 3$. We define the corresponding extracted sensing matrix $\boldsymbol{\Psi}^{(p)}$ and extracted sensor image patches $\mathbf{I}^{(p)}$ such that $\mathbf{I}^{(p)} = \boldsymbol{\Psi}^{(p)}\mathbf{L}^{(p)}$, where $\boldsymbol{\Psi}^{(p)} \in \mathbb{R}^{n_{\mathrm{s}} rq \times 3\nu q}$ and $\mathbf{I}^{(p)} \in \mathbb{R}^{n_{\mathrm{s}} rq}$. Assuming that $\mathbf{L}^{(p)}$ has a sparse representation in $\mathbf{D}$, there exists a sparse vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_{n_{\mathrm{d}}}) \in \mathbb{R}^{n_{\mathrm{d}}}$ such that $\mathbf{D}\boldsymbol{\beta} = \sum_{d=1}^{n_{\mathrm{d}}} \beta_d \mathbf{D}^{(d)}$ approximates $\mathbf{L}^{(p)}$ (i.e. $\mathbf{D}\boldsymbol{\beta} \simeq \mathbf{L}^{(p)}$). Therefore, the reconstruction of $\mathbf{L}^{(p)}$ from $\mathbf{I}^{(p)}$ given $\boldsymbol{\Psi}^{(p)}$ and $\mathbf{D}$ amounts to minimizing the following convex energy:

$$E_{\mathbf{D}}^{(p)}(\boldsymbol{\beta}) = \frac{1}{2}\,\|\boldsymbol{\Psi}^{(p)}\mathbf{D}\boldsymbol{\beta} \;-\; \mathbf{I}^{(p)}\|_2^2 \;+\; \eta\,\|\boldsymbol{\beta}\|_1\;, \qquad (8)$$

which can be efficiently solved by many existing algorithms (e.g. OMP [44], LARS [45], ISTA [46], etc.) as used in [17], [18] with different sensing matrices. Note that the proposed formalization is very generic and can be applied to various mask-based cameras and acquisition schemes. It also allows to flexibly use different BPDN solvers by considering the appropriate sensing operator $\boldsymbol{\Psi}^{(p)}$ for a given acquisition scenario.

The computation complexity of dictionary-based reconstruction algorithms is dependent on the patch size, as well as the overlapping patch distance (i.e. the pixel-wise corner-to-corner distance between two neighboring patches that overlap with each other). A larger patch size will lead to a higher computational and storage complexity. Similarly, a smaller value for patch distances increases the computational complexity. On the other hand, both of these parameters affect the reconstruction quality. The larger the patch size, the sparser the representation, leading to a higher reconstruction quality. The smaller the overlapping distance, the higher the number of pixels to be reconstructed and averaged to get the final pixel value, hence a higher image quality. As shown in [17], [18], dictionary-based methods typically produce noise-like artifact when using non-overlapping patches, specially around the sharp edges between the foreground and background. In

contrast, the use of overlapped patches is shown to be able to improve significantly the reconstruction quality (with a gain of several dB, see [18]); however it increases the computational complexity by orders of magnitude. As a result, finding suitable values for the patch size and overlapping distance that give an optimal trade-off between complexity and performance is often a tedious process. In the next section, we describe a differential-based reconstruction algorithm that aims at reducing block and noise-like artefacts resulting from inadequate patch size and overlapping distance.

B. Differential-based reconstruction with patch-based constraint

Total variation (TV) [47] is one of the most well-known regularizers in image processing. In light field processing, TV-based regularization is commonly used to regularize angular images and epipolar plane images (EPIs) for light field denoising, impainting or depth estimation [48], [49]. We introduce in this section a TV functional for 4D color light fields by taking into account the two-plane parameterization and the differential with arbitrary order.

Let us consider the following partial derivative operator applied to the light field $\mathbf{L}$ at the (discrete) spatio-angular coordinates $(\boldsymbol{i},\boldsymbol{j}) \in \mathcal{I}\times\mathcal{J}$ given by:

$$(\boldsymbol{\partial}^{\boldsymbol{\alpha}}\mathbf{L})^{\mathrm{c}}_{\boldsymbol{i},\boldsymbol{j}} = \left(\frac{\partial^{\|\boldsymbol{\alpha}\|_1}\,\mathbf{L}}{\partial x^{\alpha_1}\,\partial y^{\alpha_2}\,\partial u^{\alpha_3}\,\partial v^{\alpha_4}}\right)^{\mathrm{c}}_{\boldsymbol{i},\boldsymbol{j}} = \left(\frac{\partial^{\alpha_1}}{\partial x^{\alpha_1}}\frac{\partial^{\alpha_2}}{\partial y^{\alpha_2}}\frac{\partial^{\alpha_3}}{\partial u^{\alpha_3}}\frac{\partial^{\alpha_4}}{\partial v^{\alpha_4}}\mathbf{L}\right)^{\mathrm{c}}_{\boldsymbol{i},\boldsymbol{j}} ,$$

where $\boldsymbol{\alpha} = (\alpha_1,\alpha_2,\alpha_3,\alpha_4) \in \mathbb{N}^4$ is a non-negative integer vector, $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^4 \alpha_i$ is called the order of $\boldsymbol{\partial}^{\boldsymbol{\alpha}}$ and $\frac{\partial^{\alpha}}{\partial\,\mathrm{dir}^{\alpha}}$ denotes the α-order directional derivative with respect to the direction $\mathrm{dir} \in \mathcal{O} = \{x,y,u,v\}$. To simplify notations, let $(\nabla^K\mathbf{L})_{\boldsymbol{i},\boldsymbol{j}} = \{(\boldsymbol{\partial}^{\boldsymbol{\alpha}}\mathbf{L})^{\mathrm{c}}_{\boldsymbol{i},\boldsymbol{j}}\}_{\mathrm{c}\in\mathcal{C},\|\boldsymbol{\alpha}\|_1=K}$ denote the vector that gathers all K-order partial derivative of $\mathbf{L}$ at $(\boldsymbol{i},\boldsymbol{j})$, we then define the K-order-differential regularizer of $\mathbf{L}$ as follows:

$$\mathrm{DV}_K(\mathbf{L}) = \|(\nabla^K\mathbf{L})_{\boldsymbol{i},\boldsymbol{j}}\|_{1,2} := \sum_{\boldsymbol{i}\in\mathcal{I}}\sum_{\boldsymbol{j}\in\mathcal{J}}\sqrt{\sum_{\|\boldsymbol{\alpha}\|_1=K}\sum_{\mathrm{c}\in\mathcal{C}}\left[(\boldsymbol{\partial}^{\boldsymbol{\alpha}}\mathbf{L})^{\mathrm{c}}_{\boldsymbol{i},\boldsymbol{j}}\right]^2} ,$$

where $\|\cdot\|_{1,2}$ denotes the "mixed $\mathrm{l}_1-\mathrm{l}_2$ norm". In particular, when $K=1$, $(\nabla^1\mathbf{L})_{\boldsymbol{i},\boldsymbol{j}}$ is called the "spatio-angular gradient" computed at every position $(\boldsymbol{i},\boldsymbol{j})$ of the light field $\mathbf{L}$ and $DV_1(\mathbf{L})$ can be seen as the total variation of $\mathbf{L}$ with respect to the spatio-angular gradient operator ∇^1.

Considering the proposed regularizer, we aim at solving the following minimization problem:

$$\min_{\mathbf{L}\in\mathbb{R}^{3\nu n}} \quad \frac{1}{2}\|\boldsymbol{\Psi}\mathbf{L}-\mathbf{I}\|_2^2 + \mu\,\mathrm{DV}_K(\mathbf{L}) + \frac{\rho}{2}\|\mathbf{P}\mathbf{L}-\widehat{\boldsymbol{\mathcal{L}}}\|_2^2 , \quad (9)$$

where $\widehat{\boldsymbol{\mathcal{L}}} = \{\widehat{\mathbf{L}}^{(p)}\}_{1\le p\le \widehat{n}_{\mathrm{p}}}$ is a collection of pre-estimated light field patches, $\mathbf{P}$ denotes the extraction operator applied to $\mathbf{L}$ with respect to the patch location $\widehat{\boldsymbol{\mathcal{L}}}$, and $\mu,\rho\ge 0$ are some non-negative parameters. While the first quadratic term $\|\boldsymbol{\Psi}\mathbf{L}-\mathbf{I}\|_2^2$ measures the fidelity between the captured sensor images $\mathbf{I}$ and the projections of $\mathbf{L}$ according to the imaging matrix $\boldsymbol{\Psi}$, the second quadratic term $\|\mathbf{P}\mathbf{L}-\widehat{\boldsymbol{\mathcal{L}}}\|_2^2$ measures the similarity between the extracted patches of $\mathbf{L}$ and the given estimates $\widehat{\boldsymbol{\mathcal{L}}}$. Note that the users are free to use any convenient methods (e.g., dictionary-based or deep-learning-based methods) in order to compute $\widehat{\boldsymbol{\mathcal{L}}}$. In practice, the latter can be obtained by gathering the reconstruction results of different methods or of the same method but with different parameters. For simplicity, we only consider here patches of the same size reconstructed using dictionary-based methods.

A possible choice to solve Eqn. (9) is the proximal gradient method (also known as forward-backward splitting method [50], [46]) that requires computing $\mathrm{prox}_{\mathrm{DV}_K}$ (the proximity operator of DV_K). It is however not easy to implement $\mathrm{prox}_{\mathrm{DV}_K}$ in the case of high dimensional light fields. In this work, we consider instead the full-splitting approach [51] which allows us to design an iterative algorithm using only "simple" operations as follows:

Choose the parameters $\gamma,\tau>0$ and the initial estimates $\mathbf{L}^{(0)}\in\mathbb{R}^{3\nu n}$, $\mathbf{K}^{(0)}\in\mathbb{R}^{3\nu n 4^K}$. Then iterate, for $\ell\ge 0$:

$$\widetilde{\mathbf{L}}^{(\ell+1)} = \boldsymbol{\Psi}^*(\boldsymbol{\Psi}\mathbf{L}^{(\ell)}-\mathbf{I}) + \rho\mathbf{P}^*(\mathbf{P}\mathbf{L}^{(\ell)}-\widehat{\boldsymbol{\mathcal{L}}}) , \quad (10)$$

$$\mathbf{L}^{(\ell+1)} = \mathbf{L}^{(\ell)} - \gamma\big[\widetilde{\mathbf{L}}^{(\ell+1)} + (\nabla^K)^*\mathbf{K}^{(\ell)}\big] , \quad (11)$$

$$\widetilde{\mathbf{K}}^{(\ell+1)} = \nabla^K\big(2\mathbf{L}^{(\ell+1)}-\mathbf{L}^{(\ell)}\big) , \quad (12)$$

$$\mathbf{K}^{(\ell+1)} = \mathrm{prox}_{\tau(\mu\|\cdot\|_{1,2})^*}\big(\mathbf{K}^{(\ell)}+\tau\widetilde{\mathbf{K}}^{(\ell+1)}\big) , \quad (13)$$

in which $\mathbf{L}^{(\ell)}$ converges to a solution of Eqn. (9) if the proximal parameter condition is satisfied:

$$\gamma\left(\frac{1}{2}\big\|\boldsymbol{\Psi}^*\boldsymbol{\Psi}\big\| + \frac{\rho}{2}\big\|\mathbf{P}^*\mathbf{P}\big\| + \tau\big\|(\nabla^K)^*\nabla^K\big\|\right) < 1 .$$

Here, $\mathbf{K}^{(\ell)}$ are dual variables (in the sense of Fenchel–Rockafellar duality [52]) and $\mathbf{K}^{(0)}$ is usually initialized with 0. The functional $(\mu\|\cdot\|_{1,2})^*$ denotes the Fenchel–Rockafellar conjugate of $\mu\|\cdot\|_{1,2}$, which satisfies the Moreau identity as: $\mathrm{prox}_{\tau(\mu\|\cdot\|_{1,2})^*}(\mathbf{K}) = \mathbf{K} - \tau\mathrm{prox}_{(\mu\|\cdot\|_{1,2})/\tau}(\mathbf{K}/\tau)$.

Note that the proposed formulation can be applied not only to regularly overlapped patches, but also to patches that are non-overlapped, sparsely extracted or extracted in an irregular manner. It provides thus a flexible way to incorporate different patch-based estimators and improve the reconstruction quality. Moreover, in comparison with dictionary-based approaches, the proposed algorithm allows to obtain reconstructed results with homogeneous regions while avoiding block artifacts and noise-like artifacts that often happen when aggregating small patches as observed in [17], [18] (see [20] for more details).

VII. Experimental results

A. Datasets and acquisition setups

We evaluate the proposed algorithms on both synthetic and natural light fields. The synthetic light fields used in the experiments are collected from the MIT Media Lab archive [54] and have 5×5 views of 840×593 pixels. The natural light fields come from the Stanford Lytro Light Field Archive [22] and Kalantari's dataset [55]. These natural datasets have been captured using the Lytro Illum camera and have an angular resolution of 14×14 viewpoints for the first dataset (resp. 8×8 viewpoints for the second). However, they suffer from strong vignetting effects on peripheral views due to mechanical and optical imperfections. For that reason, we only take into account the 5×5 center angular images as reference (a.k.a ground truth) in our experiments.

A comparison between different masks patterns has been reported in [17], [18]. In the scope of this paper, we will mainly focus on the comparison between different acquisition scenarios using different light field reconstruction methods. To have a homogeneous study on different light field reconstruction approaches, we consider the camera architecture featuring one main lens and one color (coded) attenuation mask inserted

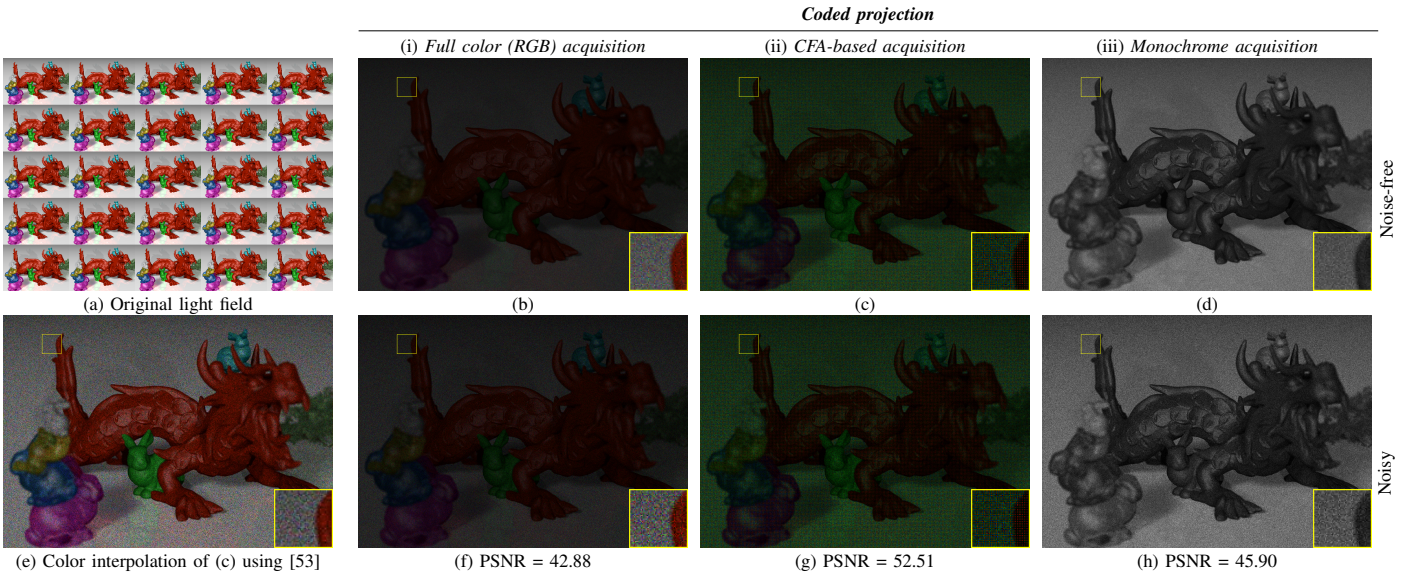


Fig. 5: **Compressed acquisition of light field with three different scenario.** The original light field is the "Dragon and bunnies" scene (5×5 views) from the MIT Archive [54]. The exposure time for each color channel of the RGB acquisition (scenario (i)) is $\frac{\triangle_t}{3}$ For each shot of a three-shot acquisition using monochrome sensor, the exposure time equals to $\frac{\triangle_t}{3}$, while it is $\triangle_t$ for a one-shot acquisition using CFA sensor. The total exposure time for each acquisition scheme remains equal to $\triangle_t$. The quality of captured images (in terms of brightness and signal-to-noise ratio) is indeed affected by the exposure time. Note that the zooms are contrast-enhanced for better visualization, while the original image intensities keep unchanged.

between the main lens and the camera sensor which can be either monochrome or CFA-based. Note that the combination of color coded mask and CFA sensor have been recently reported in [20]. It is nothing else than a particular case of our EMMC model (see Fig. 4d) and can be straightforwardly modeled using Eqn. (4). We assume that the implemented CFA is the well-known Bayer pattern [56] in all our simulations. Nevertheless, other CFA patterns such that those proposed in [57], [58] can be also considered without modification of the reconstruction algorithm.

Regarding the choice of coded masks, we use a random RGBW (stands for red-green-blue-white) pattern as the attenuation mask which is placed before the sensor, representing the same setup as described in [21], in order to provide a fair comparison with it. This kind of RGBW mask is actually a random color coded mask, similar to the one considered in [18]; the two masks differ from each other by the distributions used for generating the color patterns. Using the same RGBW attenuation mask, we consider three acquisition scenarios:

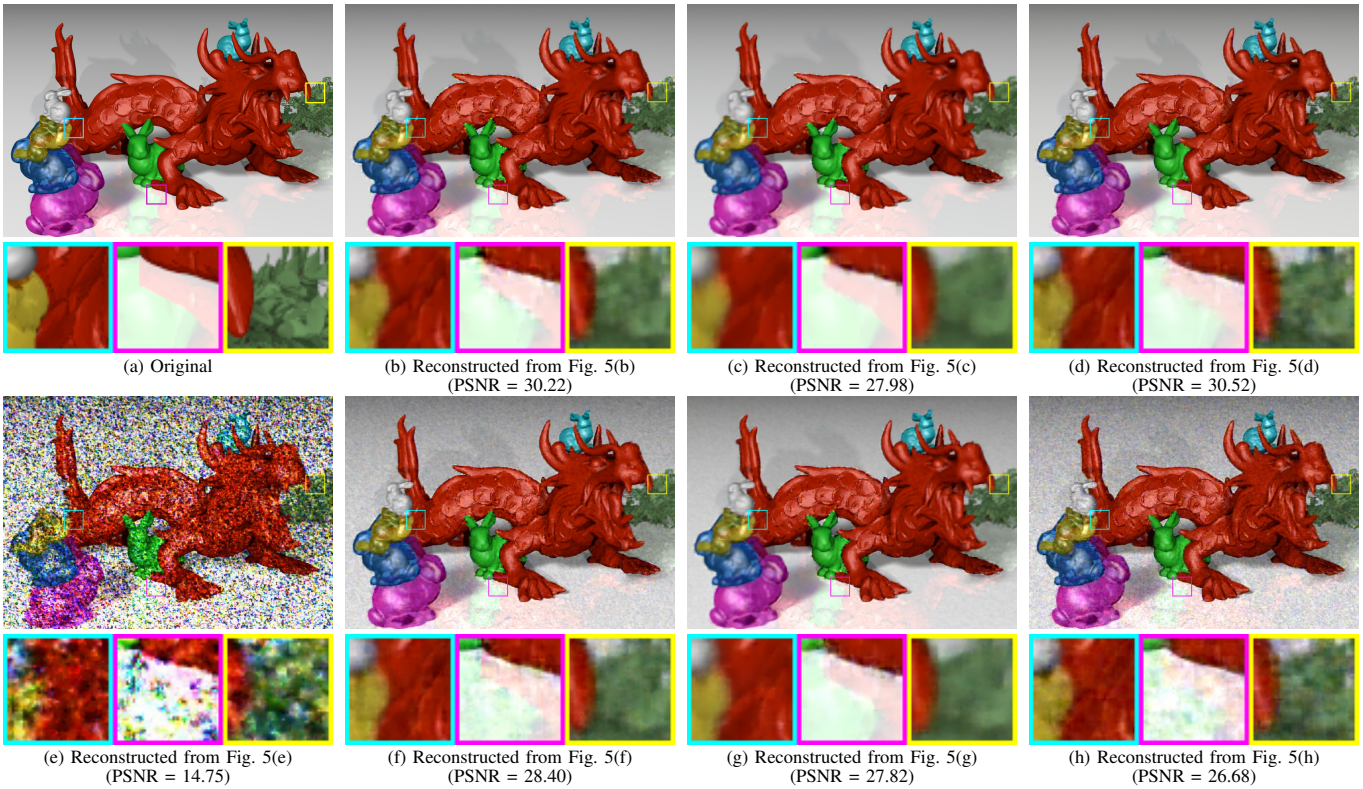
(i) color-by-color (or multi-spectral) acquisition as in [30], [17], [18], in which each coded projection is a (full color) RGB image;
(ii) acquisition with a CFA-equipped sensor as in [20], in which each coded projection is a color mosaic;
(iii) acquisition with a monochrome sensor (i.e. bare photo-sensitive sensor without CFA) as in [21], in which each coded projection is just a gray-scale image.

Examples of captured images on the sensor and reconstruction results for the three scenarios are illustrated in Figs. 1, 5 and 6. The readers are refereed to Section VII-B for further technical details of the reconstruction algorithms.

As a side note, one may recall that the first scenario corresponds to performing three acquisitions (i.e. three shots) using an equivalent multi-mask camera equipped with red, green and blue color filters as specific color masks in addition to the used RGBW mask and monochrome sensors. In contrast, the second scenario is equivalent to performing only one acquisition using an EMMC composed of the RGBW attenuation mask and a CFA mask mounted on the monochrome sensor. For the two first acquisition scenarios, the number of masks in the EMMC representation is $n_{\mathrm{m}} = 2$. The third one, which is the simplest among the three acquisition scenarios, represents the "minimal" configuration of the EMMC model featuring a coded mask and a monochrome sensor array, in which each captured image corresponds to a single-shot acquisition. For the scenarios (ii) and (iii) which require only one shot per captured coded projection, the light field reconstruction from single and multiple shots are envisaged. Moreover, acquisitions with different pixel sizes are also considered for all three scenarios.

B. Light field reconstruction

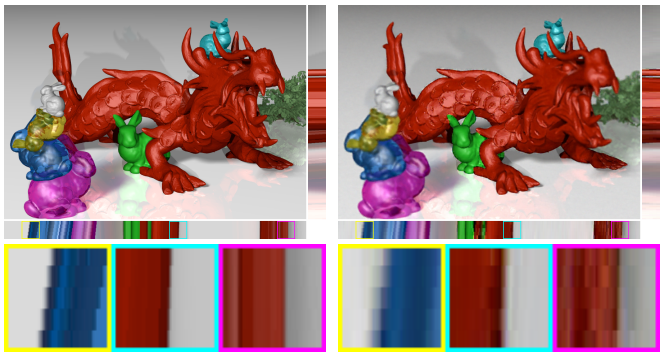
In this section, we analyze the above acquisition scenarios, encompassed by the proposed camera model, and compare the proposed differential and dictionary-based light-field reconstruction approach with a dictionary-based approach. For sake of simplicity, we consider here the first-order differential regularization (i.e. the so-called "spatio-angular" total variation). Hereafter, we refer the reconstruction algorithm that incorporates total variation and dictionary-based estimation constraint (Section VI-B) as the "differential and dictionary based algorithm" (abbreviated by "TV-Dict"). For comparison purposes, we use the ADMM algorithm [59] for the sparse decomposition task by adapting the sensing matrix in each acquisition scenario and name this dictionary-based reconstruction method the "Dict-ADMM" method. Note that the authors of [18] use the SL0 algorithm [60] to compute the sparse decomposition, resulting to an other dictionary-based method which is different to Dict-ADMM. For the latter, we consider light field patches of spatial dimension 9×9 pixels with a stride of 5 pixels. In our experiments, the TV-Dict algorithm uses Dict-ADMM results for the patch-based constraint, however, results obtained with any dictionary-based reconstruction algorithms can be used instead. Also, for sake of reference, we provide the performance achieved by the deep-learning-based method



(a) Original (b) Reconstructed from Fig. 5(b) (PSNR = 30.22) (c) Reconstructed from Fig. 5(c) (PSNR = 27.98) (d) Reconstructed from Fig. 5(d) (PSNR = 30.52)

(e) Reconstructed from Fig. 5(e) (PSNR = 14.75) (f) Reconstructed from Fig. 5(f) (PSNR = 28.40) (g) Reconstructed from Fig. 5(g) (PSNR = 27.82) (h) Reconstructed from Fig. 5(h) (PSNR = 26.68)

Fig. 6: **Visual comparison between reconstructions of the three acquisition scenarios using the proposed TV-Dict algorithm.** The top-left viewpoint and corresponding zoom-in views are extracted from the 4D light fields for comparison All results are obtained with the parameters $\mu = 2^{-8}$ and $\rho = 2^{-7}$. The three scenarios produce visually similar results when the reconstruction is performed from noise-free coded projections. In this noiseless situation, the three-shot acquisition with monochrome sensor gets the best PSNR score compared to one-shot RGB and CFA acquisitions. When there is sensor noise, the reconstruction quality drops for all scenarios. However, the one-shot CFA acquisition scheme provides very close reconstruction results for the noise-free and noisy cases.



(a) Original (b) Reconstructed from Fig. 5(g)

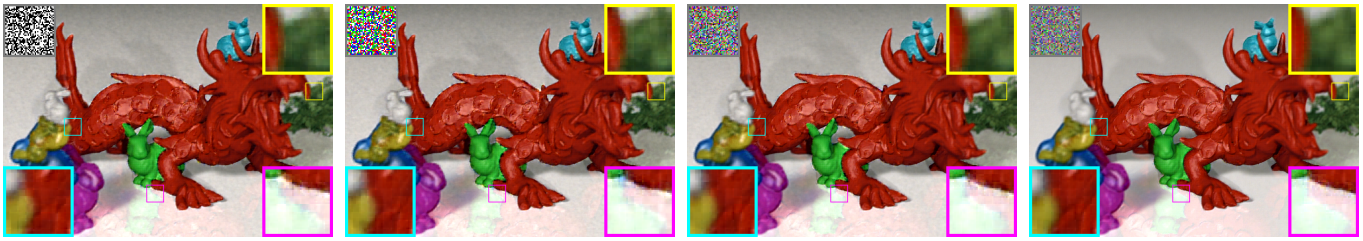
Fig. 7: **EPI visualization of original and reconstructed light fields.** The reconstruction result is obtained from the CFA acquisition scenario using the proposed TV-Dict algorithm. The reconstructed EPIs look similar to the original counterparts.

[21] whose the developed network architecture can only deal with single-shot acquisitions using monochrome sensors.

1) Results on synthetic data: In this section, we present the reconstruction results obtained with the proposed TV-Dict algorithm for the three acquisition scenarios on four test light fields collected from [54]. In this experience, we used a dictionary trained on randomly selected patches which were extracted from the scenes of the same data set (excluding the test set), using the K-SVD algorithm [61][1].

a) Exposure time impact analysis: We note that all the acquisitions in this experiment are performed with the same amount of exposure time to conduct a fair comparison between them. Here, we consider three-shot acquisitions when the monochrome sensor is used, since only one monochrome captured image does not allow to recover color information as depicted in [62]. More precisely, given a total time $\triangle_t$, the acquisition time of each color channel of full RGB images is $\frac{\triangle_t}{3}$, the latter is also the time for each shot of a three-shot acquisition using monochrome sensor, while the exposure time for a single-shot acquisition using CFA sensor remains equal to $\triangle_t$. Acquired images in these above settings are illustrated in Fig. 5 showing the effect of the exposure time on the image quality in terms of brightness and signal-to-noise ratio (SNR). The RGB acquisition (c.f. the second column of Fig. 5) produces indeed very dark coded projections when comparing to the two other scenarios (due to short exposure time and color filtering). Consequently, its sensor image is much more noisier (with PSNR = 42.88 dB – the lowest among the three scenario). The PSNR is computed taking the noiseless acquisition as a reference, it hence measures the quality of the acquired measurements. Having the same exposure time, the acquisition using monochrome sensor provides brighter coded projections (c.f. the fourth column of 5), implying better signal-to-noise ratio (PSNR = 45.90 dB) since the sensor can integrate photons of all the light wavelengths. As expected, the acquisition using CFA sensor (c.f. the third column of Fig. 5), which benefits from longer exposure time, tends to achieve the best PSNR score (over 50 dB) while maintaining the same overall brightness level as obtained with monochrome sensor (see its color-interpolated version shown in Fig. 5e for better visual evaluation instead of using the original one). Note

[1]The K-SVD toolbox is available at http://www.cs.technion.ac.il/~ronrubin/software.html



(a) RGBW (reconstructed from Fig. 5(g) with Dict-ADMM – PSNR = 27.43) (b) Binary (PSNR = 27.43) (c) Uniform (PSNR = 26.95) (d) Gaussian (PSNR = 26.28)

Fig. 8: **Reconstruction results obtained with different coded masks and a CFA-based sensor using the Dict-ADMM algorithm**. From left to right: (a) the RGBW pattern (as in [21]), (b) the random binary pattern (following a Bernoulli's distribution of probability 0.5), (c) the "uniform color" pattern (each color component of the mask follows a uniform law in $[0, 1]$) and (d) the "Gaussian color" pattern (each color component of the mask follows a Gaussian distribution centered at 0.5 with the standard variation 0.2 and values greater than 1 or smaller than 0 are clipped).

that this interpolation is obtained with the classical gradient-correction-based method [53]. Although it looks similar to the captured image obtained by a full (RGB) color acquisition scheme (see Fig. 5b), one can remark that the demosaicing tends to oversmooth color transition and thus fails to recover high frequency color information.

Corresponding reconstruction from the coded projections of the three acquisition scenarios is shown in Fig. 6. All the results are obtained with the proposed variational algorithm (TV-Dict) using the parameters $\mu = 2^{-8}$ and $\rho = 2^{-7}$. For the comparison purposes, we present the results gathered from two situations: with and without sensor noises. When no noise is added, the three scenarios produce visually similar results (c.f. the first row of Fig. 6) and they are very close to the original light field (see Fig. 6a). In this situation, the reconstruction from three coded projections acquired with monochrome sensor (see Fig. 6d) is slightly better both in terms of visual and quantitative evaluations when compared to those obtained from only one coded projection in the case of CFA or RGB acquisition. It is mainly due to the higher number of recorded measurements (three-fold over one-shot CFA acquisition) and higher amount of incident light (three-fold over one RGB captured image).

Among the three scenarios with the same total exposure time, the one-shot CFA acquisition, which achieves the best SNR of (noisy) sensor images while having the less recorded samples, produces decent reconstruction results: slightly inferior PSNR (27.82 compared to 28.40 obtained with one RGB acquisition) and visually pleasant restoration of fine details (c.f. Fig 6g). In contrast, the results obtained from the two other schemes look much more noisy (see Figs. 6f and 6h) although the number of recorded samples is three times higher. One can see that the overall reconstruction quality depends strongly on how noisy the acquired images are and how many images to be acquired. In our opinion, considering the three acquisition scenarios, the one using CFA sensor provides the best trade-off between the SNR and the number of measurements. Moreover, it also allows to retain some color information thanks to the CFA pattern (compared to the lost of all color information in the case of monochrome sensor).

As a side note, we may inform the readers that the reconstruction from pre-demosaiced coded projection (i.e. the color-interpolated version of the CFA acquisition, c.f. Fig 5e) is depicted in Fig. 6e for illustration purposes. The obtained result, which suffers from heavy noise-like artifacts, can be seen as a reconstruction from very noisy RGB acquisitions. Similar results with a different coded mask pattern are reported in [20]. Also, in order to demonstrate the performance of the TV-Dict algorithm in terms of parallax reconstruction, we show in Fig. 7b the epipolar plane image (EPI) visualization of the reconstruction result from CFA acquisition. Although the reconstructed EPIs is slightly noisy compared to the original one (see Fig. 7a), we can easily remark that most slopes have been well restored.

b) Noise impact analysis: In order analyse the behavior of our approach in presence of noise, the coded projections are corrupted by random noises at the sensor level (i.e. the digitization stage). For the sensor noise model, we consider a mixed Poisson-Gaussian noise combined with a quantization noise to cope with the quantum nature of the light as well as the on-sensor electronic fluctuations and the analog-to-digital conversion (ADC) procedure. Accordingly, the sensor image can be expressed as:

$$\mathbf{I}_{\text{sensor}} = g_{\text{ADC}} \left[\mathcal{P}(\triangle_t \mathbf{I}) + \mathbf{n}_{\text{read}} \right] + \mathbf{n}_{\text{quan}} \ , \qquad (14)$$

where $\mathbf{n}_{\text{read}} \sim \mathcal{N}(0, \sigma^2_{\text{read}})$ is a Gaussian variable representing the electronic read-out noise, $\mathbf{n}_{\text{quan}} \sim \mathcal{U}([0, 1])$ follows the uniform distribution on the interval $[0, 1]$ and models the quantization at the ADC level, $g_{\text{ADC}} > 0$ denotes the ADC gain (also called "camera gain"), $\triangle_t$ is the exposure time and $\mathcal{P}(\triangle_t \mathbf{I})$ is the Poisson variable modeling the photon shot noise with respect to the amount of incident light $\mathbf{I}$ and the exposure time $\triangle_t$. In the experiments reported below, the

		(i)		(ii)		(iii)	
		Dict-ADMM	TV-Dict	Dict-ADMM	TV-Dict	Dict-ADMM	TV-Dict
Noise-free	Dragon	29.46	30.22	27.62	27.98	30.01	30.52
	Dice	28.67	28.93	25.74	25.78	28.92	29.29
	Fish	27.47	27.86	25.48	25.53	28.01	28.33
	Messerschmitt	32.42	32.51	29.96	29.89	32.17	32.93
Noisy	Dragon	28.00	28.40	27.43	27.82	26.35	26.68
	Dice	26.75	27.02	25.42	25.54	25.05	25.33
	Fish	27.07	27.47	25.07	25.20	26.18	26.42
	Messerschmitt	29.98	30.67	29.64	29.75	28.46	29.21

TABLE II: **PSNR obtained with the three acquisition schemes, (i) full color (RGB), (ii) CFA-based and (iii) monochrome acquisitions, and three reconstruction methods, with both noise-free and noisy acquisitions**. The light field used in the test is a synthetic light field from the MIT dataset [54]. The regularization parameter μ has been set to 0.01 and 0.05 in the noise-less and noisy cases respectively.

camera gain is set as $g_{\mathrm{ADC}} = 0.8$ (i.e. each generated photo-electron corresponds to 0.8 digital unit of pixel intensity in the output sensor image), the standard deviation of the read-out noise $\sigma_{\mathrm{read}} = 2.5$ (electrons RMS) and the total exposure time $\triangle_t = \frac{1}{20}$ second.

By considering this noise model, it is also possible to replace the data fidelity term $\frac{1}{2}\|\mathbf{\Psi L} - \mathbf{I}\|_2^2$ (which is adapted to an additive Gaussian noise) in Eqn. (8) and (9) by a Bayesian-based Poisson-Gaussian data term as suggested in [63], [64], [65]. However, in the context of photography, when the illumination and the exposure conditions are sufficient, the photon shot noise involved in the imaging process can be approximated by a Gaussian noise. For that reason, we propose to simply substitute the noise-free coded projection $\mathbf{I}$ by its noisy version $\widehat{\mathbf{I}} = \frac{1}{\triangle_t g_{\mathrm{ADC}}}\mathbf{I}_{\mathrm{sensor}}$ in Eqn. (8) and (9) to scale the captured image $\mathbf{I}_{\mathrm{sensor}}$ into the same intensity space of the original light field $\mathbf{L}$ when performing the reconstruction from noise-corrupted coded projections.

We provide in Table II the PSNR values obtained on four test light fields, considering both noise-free and noisy acquisitions. When sensor noise is taken into account, one can remark that the reconstruction quality decreases drastically for all the three scenarios, regardless the choice of reconstruction algorithm (see Table II for more details). In this study, we compare the TV-Dict method with the ADMM-based implementation of the dictionary-based reconstruction (Dict-ADMM) for all the three acquisition scenarios. Based on reported results in Table II, the former method obtains higher PNSR values in most cases, compared to the latter. In our opinion, it is due to the fact that the differential regularization which favors homogeneous reconstructed regions, and thus reducing the noise and eventually block artifacts presented in patch-based reconstruction (see [20]). This also demonstrates the effectiveness of our regularization-based method on the improvement of light field reconstruction quality in different acquisition scenarios.

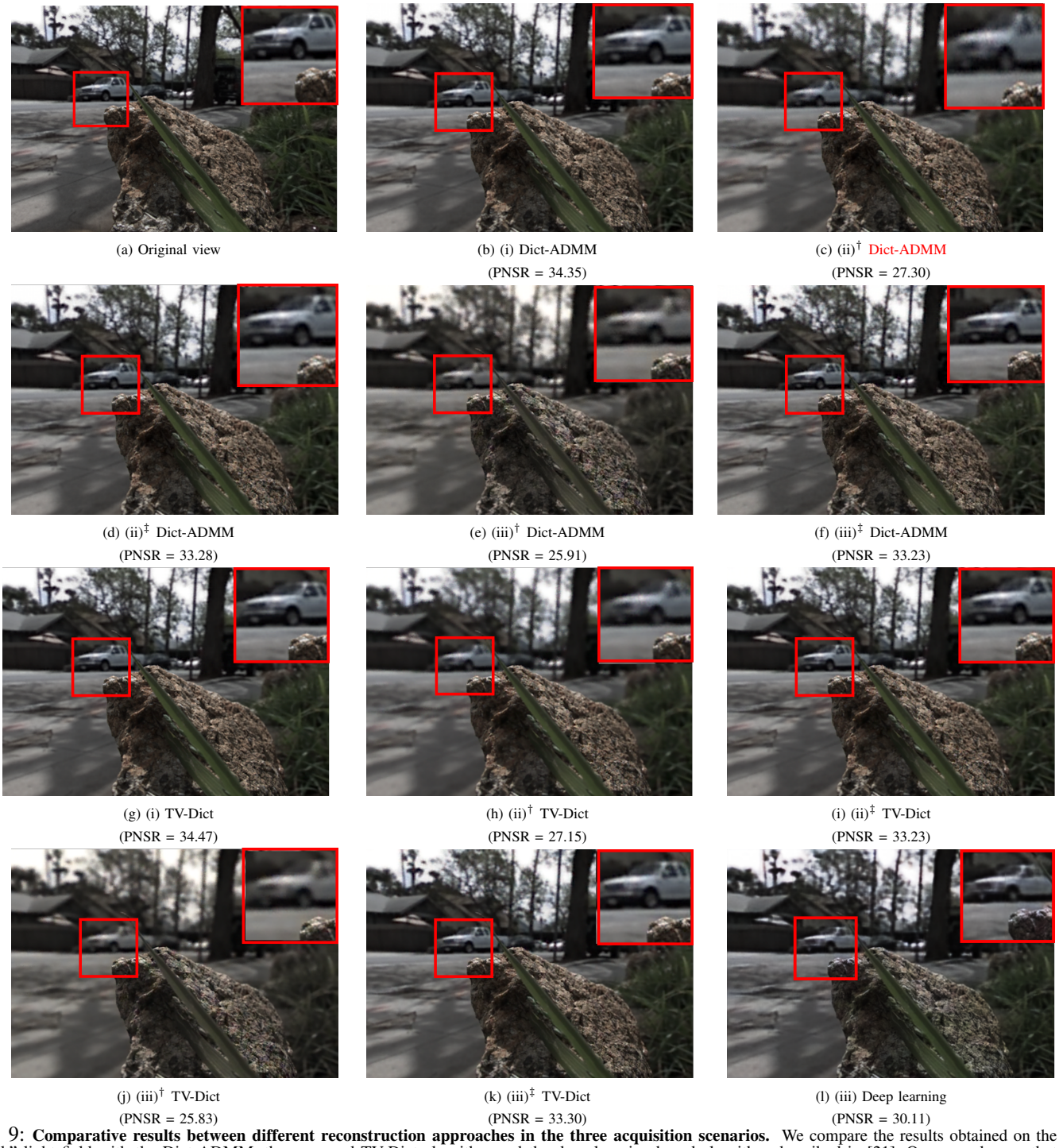
While there are several published works that already report the comparison between different types of coded masks (see [17] and [18]), none of them actually considers a realistic sensor (e.g. CFA-equipped sensors) neither a realistic noise model (which takes into account the non-stationary nature of the noise in real life). Here, we illustrate in Fig. 8 the reconstruction results obtained with various coded masks, using the Dict-ADMM algorithm for the CFA-based acquisition scenario (ii) in the presence of sensor noise as in Eqn. (14) with the same parameters as above (i.e. $\sigma_{read} = 2.5, g_{\mathrm{ADC}} = 0.8$ and $\triangle_t = \frac{1}{20}$). In this experiment, we consider the following mask patterns: the RGBW pattern, the random binary pattern (following a Bernoulli's distribution of probability 0.5), the "uniform color" pattern (i.e. each color component of the mask follows a uniform law in $[0, 1]$) and the "Gaussian color" pattern (where each color component of the mask follows a Gaussian distribution centered at 0.5 with the standard variation 0.2 and values greater than 1 or smaller than 0 are clipped). We can easily remark that all these masks have the same light transmission rate which is approximately equal to 50%. The results depicted in Fig. 8 show that best PSNR performances are obtained using the RGBW mask and the binary mask (see Fig. 8a and 8b), compared to the uniform and Gaussian masks (see Fig. 8c and 8d), although these latter masks give better randomness in terms of mask value variety. Based on this observation, we may think that the level of sensor noise has a significant impact on the choice of the implemented coded mask to achieve a target quality for light field reconstruction. Indeed, this can be an interesting subject for future studies.

2) Results on real light fields: In this section, we compare the results obtained, with different reconstruction methods for the three acquisition scenarios, on real light field data captured using a Lytro Illum camera. Unlike the experiments reported in Table II, no simulation noise has been added in this experiment, since the tested light fields already contain noise and therefore the characteristics of sensor noise cannot be properly controlled when dealing with different acquisition scenarios and exposure times. Instead of studying the impact of noise on the reconstruction quality, we focus here on a scenario-by-scenario analysis as follows:

(i) Multi-spectral (or color per color) acquisition: Table III shows that scenario (i) gives a better reconstruction quality with both the dictionary-based and TV-based approaches. However, its compression ratio is three-time-lower (i.e. it captures three times more samples) than the ones of scenarios (ii) and (iii). In addition, this scenario (considered in [16], [17], [33], [18]) is not appropriate for standard consumer cameras due to major modifications of the camera architecture (e.g. using prisms to separate light colors and using a monochrome sensor to capture each color component).

(ii) With a CFA mask (sampled color acquisition): Table III shows that this dual-mask design (using a coded mask and a CFA mask) combined with the proposed joint demosaicing-reconstruction approach, works well and gives decent results. The above solution (the dual-mask design combined with the TV-Dict algorithm) is compatible with existing sensors with in-built CFAs and allows reconstruction from multiple shot acquisitions, which is not the case of [21]. In fact, the neural network developed in [21] only allows the reconstruction from single coded projections and can not deal with multiple-shot acquisitions without changing the network architecture and retraining the modified network. In addition, it is sensitive to the distribution of mask colors since it needs to be trained for each specific mask pattern (or color distribution) in order to achieve its best performance (see Appendix A of [21] for more details). The TV-Dict approach, in contrast, does not possess this restriction and can handle both single and multiple shots as well as various coded mask patterns (including monochrome and color masks).

(iii) Using a monochrome sensor with one or multiple shots: This scenario, referred to as scenario (iii) in Table III, with one shot, records the same number of measurements as scenario (ii) but three times less than scenario (i). We can observe that in scenario (iii), all tested algorithms, including the deep learning approach (see Fig. 9f), often fail to correctly reconstruct colors of the original light fields from one single-shot acquisition (i.e. from only one coded projection) due to the lack of color information. In fact, the reconstructed colours (from one-shot acquisitions with the monochrome sensor) are less vivid than in the original images. The color can be better recovered by increasing the number of shots (see Fig. 9e and 9f respectively), implying longer total exposure times. Two acquisitions, i.e. two shots, are required when using a monochrome sensor for a good recovery of both colour and parallax. Table III (columns (iii)‡) shows the PSNR values obtained with the different algorithms (Dict-ADMM and TV-Dict) using two shots. The results demonstrate that the combination of a global and local reconstruction integrating a TV-based and a dictionary-based regularization (i.e. the TV-Dict algorithm) improves the results for the three acquisition scenarios.



(a) Original view

(b) (i) Dict-ADMM (PNSR = 34.35)

(c) (ii)† Dict-ADMM (PNSR = 27.30)

(d) (ii)‡ Dict-ADMM (PNSR = 33.28)

(e) (iii)† Dict-ADMM (PNSR = 25.91)

(f) (iii)‡ Dict-ADMM (PNSR = 33.23)

(g) (i) TV-Dict (PNSR = 34.47)

(h) (ii)† TV-Dict (PNSR = 27.15)

(i) (ii)‡ TV-Dict (PNSR = 33.23)

(j) (iii)† TV-Dict (PNSR = 25.83)

(k) (iii)‡ TV-Dict (PNSR = 33.30)

(l) (iii) Deep learning (PNSR = 30.11)

Fig. 9: **Comparative results between different reconstruction approaches in the three acquisition scenarios.** We compare the results obtained on the "Rock" light field with the Dict-ADMM, the proposed TV-Dict algorithm and the deep-learning-based algorithm described in [21]. One can observe that single-shot acquisitions with monochrome sensor do not allow correct recovery of color information when comparing to single-shot acquisition with CFA sensor. In our opinion, regardless of reconstruction methods, multiple-shot acquisition is necessary to correctly recover the colors when using monochrome sensor.

C. Joint light field reconstruction and spatial super-resolution

In this section, we further show that the EMMC model also covers the case where the pixel size $\triangle_p$ varies. As explained in Section V-B, in the proposed model of Eqn.(6), $\boldsymbol{H}^{(s)} \in \mathbb{R}^{rn \times n}$ is a matrix representing the integration (summation) of incident light rays on the same sensor pixels for the s-th acquisition, where $r = (\frac{\triangle_x}{\triangle_p})^2$ is the squared ratio between the spatial sampling step and the pixel width. In the experiments reported in Table IV, the pixel size is doubled in both the horizontal and vertical dimensions so that one pixel integrates the illumination of four light rays (corresponding to $r = \frac{1}{4}$). The problem in this case becomes a joint problem of reconstruction and spatial super-resolution. Table IV shows the PSNR obtained when reconstructing the light field with a spatial magnification factor of $r^{-\frac{1}{2}} = 2$ in both the horizontal and vertical dimensions, with the three acquisition scenarios. Note that one RGB (full color) acquisition with $r = \frac{1}{4}$ corresponds to a sampling rate equal to $\frac{1}{4\nu}$ (i.e. $\frac{1}{100}$ in the tests reported below), while the sampling rate of a one-shot acquisition using monochrome sensor as in [21] is $\frac{1}{3\nu}$ (i.e. $\frac{1}{75}$ in the tests below). One can see in Fig.

	Dictionary-based Dict-ADMM							Differential and dictionary-based (TV-Dict)							Learning-based [21]
	(i)	(ii)†	(ii)‡	(ii)†‡	(iii)†	(iii)‡	(iii)†‡	(i)	(ii)†	(ii)‡	(ii)†‡	(iii)†	(iii)‡	(iii)†‡	(iii)†
Seahorse	34.57	29.88	33.67	34.79	28.98	33.62	35.36	35.05	29.92	33.16	35.26	29.12	34.08	36.05	32.36
Rock	34.35	27.30	33.28	34.58	25.91	33.23	35.38	34.47	27.15	33.23	34.69	25.83	33.30	35.62	30.11
Cars	33.02	27.21	31.94	33.24	25.78	31.63	33.74	33.16	27.12	31.98	33.37	25.78	31.75	33.96	29.88
Orchid	33.73	28.42	32.88	33.91	27.25	32.57	34.78	33.98	28.41	33.00	34.16	27.28	32.77	34.90	30.99
White Rose	34.36	28.31	33.24	34.54	27.41	33.15	35.25	34.54	28.23	33.29	34.71	27.42	33.29	35.54	31.84
Tulip*	42.40	39.81	41.74	42.58	36.95	41.70	42.94	44.52	40.83	43.03	44.47	37.71	43.91	45.61	38.26
Buttercup*	33.41	29.71	32.52	33.59	28.06	32.22	34.06	33.60	29.64	32.67	33.76	28.06	32.37	34.28	29.98

TABLE III: PSNR comparison between the differential-based algorithms and dictionary-based and learning-based methods with different acquisition scenarios. (i) color-by-color acquisition corresponding to three shots using color filters with monochrome sensors; (ii) acquisition with CFA-built-in sensor and joint light field reconstruction and demosaicing; and (iii) acquisition without CFA (one shot with monochrome sensors). †, ‡ and †‡ denote reconstruction results from one-shot two-shot and three-shot acquisitions respectively. The two-shot acquisitions imply two times the number of samples compared with [21]. Light fields indicated with a "*" come from the Stanford Lytro data set [22] while the others come from the data set of [55]

.

		(i) Full RGB acquisition				(ii) Acquisition using CFA-sensor				(iii) Acquisition using monochrome sensor			
		$n_s = 3$	$n_s = 6$	$n_s = 9$	$n_s = 12$	$n_s = 1$	$n_s = 2$	$n_s = 3$	$n_s = 4$	$n_s = 1$	$n_s = 2$	$n_s = 3$	$n_s = 4$
Dict-ADMM	Seahorse	29.14	32.73	34.51	35.62	26.41	29.17	30.67	31.70	26.00	29.11	30.79	31.94
	Rock	26.38	31.98	34.21	35.53	23.19	26.91	29.23	30.81	22.88	26.73	29.30	30.93
	Cars	26.29	30.91	33.02	34.26	23.07	26.27	28.14	29.37	22.43	26.33	28.31	29.76
	Orchid	27.73	31.92	33.89	35.19	24.69	27.71	29.28	30.36	24.09	27.61	29.46	30.73
	White Rose	27.67	32.29	34.44	35.70	24.48	28.01	29.85	31.17	24.20	27.85	30.01	31.37
	Tulip*	39.42	41.58	42.64	43.24	36.39	37.53	38.23	38.69	31.47	37.75	39.36	40.21
	Buttercup*	28.88	31.62	33.31	34.56	26.48	28.00	28.88	29.62	24.32	28.19	29.54	30.44
TV-Dict	Seahorse	29.25	33.05	35.02	36.28	26.44	29.23	30.79	31.86	26.21	29.27	31.03	32.25
	Rock	26.42	32.06	34.39	35.79	23.14	26.85	29.19	30.80	22.86	26.68	29.29	30.97
	Cars	26.33	30.98	33.16	34.46	23.07	26.27	28.16	29.42	22.49	26.35	28.38	29.83
	Orchid	27.78	32.04	34.09	35.48	24.70	27.72	29.32	30.43	24.22	27.66	29.34	30.84
	White Rose	27.72	32.41	34.65	35.99	24.48	27.99	29.87	31.23	24.31	27.88	30.07	31.48
	Tulip*	40.14	43.41	44.99	45.99	37.25	38.32	39.00	39.45	31.64	38.80	40.91	42.06
	Buttercup*	28.94	31.71	33.46	34.75	26.44	28.00	28.90	29.66	24.29	28.22	29.61	30.53

TABLE IV: Super-resolution results with the Dict-ADMM method (top part of the table) and TV-Dict (bottom part of the table) based reconstruction methods (with $r = \frac{1}{4}$). n_s denotes the number of coded projections, e.g. $n_s = 3$ in the full RGB acquisition scenario means one coded projection for each color channel. Light fields indicated with a "*" come from the Stanford Lytro data set [22] while the others come from the data set of [55].

10 that, despite a higher compression factor, the model and proposed method with the full color acquisition scenario allow us to obtain a better color reconstruction quality compared with the deep learning approach. The deep learning approach of [21] assuming the use of monochrome sensors instead of CFA-based sensors, as mentioned above, cannot accurately reconstruct the image color, while the full colour acquisition scenario, despite the lower number of measurements, allows us to well recover color information. When using monochrome sensor with $r = \frac{1}{4}$, the sampling rate of one-shot acquisitions is even lower, i.e. $\frac{1}{12\nu}$ (corresponds to $\frac{1}{300}$ for color light fields with 5×5 views). One can observe the PSNR values significantly improve when increasing the number of shots. In fact, for 4 low-resolution shots with monochrome sensor, the reconstruction result reach a PSNR quality that is comparable to the one of [21] (slightly inferior) with the same number of measurements, but having better color reconstruction. In our opinion, not only the number of samples but also the way they are selected can have a huge impact on the reconstruction quality.

Our algorithm is implemented and tested on Matlab R2018b under Linux Ubuntu 18.04. All the experiments are done on a Dell Latitude 7490 featuring Intel i7 CPU with 4 cores. In terms of processing time, the reconstruction of a $9\times9\times5\times5\times3$ light field patch for 10000 iterations takes 7.5 seconds using the dictionary-based reconstruction algorithm (Dict-ADMM). The use of the differential-based reconstruction step requires 4.5 extra seconds in addition to 7.5 seconds for the dictionary-based reconstruction step.

VIII. Conclusion

We have presented a unifying camera model for compressed acquisition of Light Fields using coded masks. The proposed equivalent multi-mask camera model allows a flexible configuration of a variety of acquisition schemes. Considering the CFA pattern present in sensors as a particular mask of the proposed model led us to introduce a joint demosaicing and reconstruction method using a TV or a dictionary-based regularization. Compared with a state of the art deep learning approach, the proposed model and reconstruction method offers the possibility of dealing with multiple-shot acquisitions without changing the network architecture and retraining the network. The EMMC model in addition supports the possibility of further increasing the pixel size, thus increasing the compression rate. The reconstruction algorithm in this case jointly performs spatial super-resolution and parallax reconstruction of original light fields. Future work will be dedicated to extending the proposed model in order to take into account the diffraction effect that can occur in real high-resolution mask-based hardware imaging systems.

Acknowledgments

We thank Dr Ofir Nabati for providing the results obtained with [21], that are used for the comparison in this paper.

References

[1] M. Levoy and P. Hanrahan, "Light Field Rendering," in Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, ser. SIGGRAPH. ACM, 1996, pp. 31–42.

[2] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The Lumigraph," in Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, ser. SIGGRAPH, 1996, pp. 43–54.

[3] W.-C. Chen, J.-Y. Bouguet, M. H. Chu, and R. Grzeszczuk, "Light Field Mapping: Efficient Representation and Hardware Rendering of Surface Light Fields," ACM Transactions on Graphics, vol. 21, no. 3, pp. 447–456, Jul 2002.

[4] R. Ng, "Light Field Photography," PhD dissertation, Stanford University, Department of Computer Science, 2006.

[5] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled Photography: Mask Enhanced Cameras for Heterodyned Light Fields and Coded Aperture Refocusing," ACM Transactions on Graphics, vol. 26, no. 3, Jul 2007.

(a) Original view

(b) CFA sensor with 1 shot (PNSR = 24.70)

(c) (CFA sensor with 2 shots (PNSR = 27.72)

(d) Full RGB acquisition with 1 shot (PNSR = 27.78)

(e) Monochrome sensor with 1 shot (PNSR = 24.22)

(f) Monochrome sensor with 2 shots (PNSR = 27.66)

(g) Deep learning (PNSR = 30.99)

(h) Monochrome sensor with 3 shots (PNSR = 29.34)

(i) Monochrome sensor with 4 shots. (PNSR = 30.84)

Fig. 10: **Comparative results of super-resolution between different acquisition scenarios.** Super-resolution results obtained on "Orchid" using the proposed TV-Dict algorithm ($\mu = 2^{-8}$ and $\rho = 2^{-10}$) for the three situations: (i) full RGB acquisition, (ii) acquisition with CFA-sensor and (iii) acquisition with monochrome sensor with the super-resolution factor $r = \frac{1}{4}$. We compare with the deep-learning-based method [21] for a standard single-shot acquisition using monochrome sensor (i.e. $r = 1$).

[6] 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, vol. 8500, Oct 2012.

[7] Z. Xu, J. Ke, and E. Y. Lam, "High-resolution Lightfield Photography Using Two Masks," Optics Express, vol. 20, no. 10, pp. 10 971–10 983, May 2012.

[8] S. Lee, C. Jang, S. Moon, J. Cho, and B. Lee, "Additive Light Field Displays: Realization of Augmented Reality with Holographic Optical Elements," ACM Transactions on Graphics, vol. 35, no. 4, pp. 60:1–60:13, Jul 2016.

[9] G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar, "Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting," ACM Transactions on Graphics (SIGGRAPH), vol. 31, no. 4, pp. 1–11, Aug. 2012.

[10] T. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in IEEE International Conference on Computer Vision (ICCV), Dec 2015, pp. 3487–3495.

[11] E. H. Adelson and J. R. Bergen, "The Plenoptic Function and the Elements of Early Vision," Computational Models of Visual Processing, MIT Press, pp. 3–20, 1991.

[12] J. C. Yang, M. Everett, C. Buehler, and L. McMillan, "A Real-Time Distributed Light Field Camera," in Eurographics Workshop on Rendering (EGSR), 2002, pp. 77–86.

[13] 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, vol. 24, no. 3, pp. 765–776, Jul 2005.

[14] E. H. Adelson and J. Y. A. Wang, "Single Lens Stereo with a Plenoptic Camera," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 14, no. 2, pp. 99–106, Feb 1992.

[15] 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.

[16] 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.

[17] 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), vol. 32, no. 4, pp. 46:1–46:12, 2013.

[18] 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.

[19] L. Mignard-Debise, J. Restrepo, and I. Ihrke, "A Unifying First-Order Model for Light-Field Cameras: The Equivalent Camera Array," IEEE Transactions on Computational Imaging, vol. 3, no. 4, pp. 798–810, Dec 2017.

[20] H.-N. Nguyen and C. Guillemot, "Color and Angular Reconstruction of Light Fields from Incomplete-Color Coded Projections," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020.

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

[22] 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.

[23] C. Zhou and S. K. Nayar, "Computational Cameras: Convergence of Optics and Processing," IEEE Transactions on Image Processing, vol. 20, no. 12, pp. 3322–3340, Dec 2011.

[24] G. Lippman, "La photographie intégrale," Comptes-Rendus, Academie des

Sciences, 1908.
[25] C.-K. Liang and R. Ramamoorthi, "A light transport framework for lenslet light field cameras." ACM Transactions on Graphics (TOG), vol. 34, no. 2, pp. 1–19, 2015.
[26] L.-Y. Wei, C. Liang, G. Myhre, C. Pitts, and K. Akeley, "Improving light field camera sample design with irregularity and aberration," ACM Transactions on Graphics (TOG), vol. 34, no. 4, pp. 1–11, 2015.
[27] 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, vol. 27, no. 3, pp. 55:1–55:10, Aug 2008.
[28] 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.
[29] 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.
[30] 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.
[31] 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.
[32] 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.
[33] 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.
[34] 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.
[35] N. Pegard, H. Liu, N. Antipa, M. Gerlock, H. Adesnik, and L. Waller, "Compressive light-field microscopy for 3d neural activity recording," Optica, vol. 3, no. 5, pp. 517–524, May 2016.
[36] K. Venkataraman, D. Lelescu, J. Duparre, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar, "Picam: An Ultra-thin High Performance Monolithic Camera Array," ACM Transactions on Graphics, vol. 32, no. 3, p. 166, 2013.
[37] J. Iseringhausen, B. Goldlücke, N. Pesheva, S. Iliev, A. Wender, and M. B. H. M. Fuchs, "4d imaging through spray-on optics," ACM Transactions on Graphics (TOG), vol. 36, no. 4, pp. 35:1–35:11, 2017.
[38] M. Hirsch, S. Sivaramakrishnan, S. Jayasuriya, A. Wang, A. Molnar, R. Raskar, and G. Wetzstein, "A switchable light field camera architecture with angle sensitive pixels and dictionary-based sparse coding," in International Conference on Computational Photography (ICCP), 2014, pp. 1–10.
[39] C. K. Liang, Y.-C. Shih, and H. Chen, "Light field analysis for modeling image formation," IEEE Transactions on Image Processing, vol. 20, no. 2, pp. 446–460, Feb 2011.
[40] 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.
[41] A. Agrawal, A. Veeraraghavan, and R. Raskar, "Reinterpretable Imager: Towards Variable Post-Capture Space, Angle and Time Resolution in Photography," Computer Graphics Forum, vol. 29, no. 2, pp. 763–772, 2010.
[42] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," IEEE Signal Processing Magazine, vol. 25, no. 2, pp. 21–30, Mar 2008.
[43] S. Chen, D. Donoho, and M. Saunders, "Atomic Decomposition by Basis Pursuit," SIAM Journal on Scientific Computing, vol. 20, no. 1, pp. 33–61, 1998.
[44] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in 27th Asilomar Conference on Signals, Systems and Computers, vol. 1, 1993, pp. 40–44.
[45] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," Annals of Statistics, vol. 32, no. 2, pp. 407–499, 2004.
[46] A. Beck and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," SIAM Journal on Imaging Sciences, vol. 2, no. 1, pp. 183–202, 2009.
[47] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear Total Variation Based Noise Removal Algorithms," Physica D: Nonlinear Phenomena, vol. 60, no. 1, pp. 259–268, 1992.
[48] B. Goldluecke and S. Wanner, "The variational structure of disparity and regularization of 4d light fields," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1003–1010.
[49] N. B. Monteiro, J. P. Barreto, and J. Gaspar, "Dense lightfield disparity estimation using total variation regularization," in Image Analysis and Recognition, A. Campilho and F. Karray, Eds. Springer International Publishing, 2016, pp. 462–469.
[50] P. Combettes and V. Wajs, "Signal Recovery by Proximal Forward-Backward Splitting," Multiscale Modeling & Simulation, vol. 4, no. 4, pp. 1168–1200, 2005.
[51] L. Condat, "A Generic Proximal Algorithm for Convex Optimization–Application to Total Variation Minimization," IEEE Signal Processing Letters, vol. 21, no. 8, pp. 985–989, 2014.
[52] H. H. Bauschke and P. L. Combettes, Fenchel–Rockafellar Duality. New York, NY: Springer New York, 2011.
[53] H. S. Malvar, L.-W. He, and R. Cutler, "High-quality linear interpolation for demosaicing of Bayer-patterned color images," in International Conference on Acoustics, Speech, and Signal Processing, vol. 3, 2004, pp. iii–485.
[54] G. Wetzstein, "Synthetic light field archive," http://web.media.mit.edu/~gordonw/SyntheticLightFields/index.php, 2013, online; used in Compressive Light Field Photography Siggraph 2013 project.
[55] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based View Synthesis for Light Field Cameras," ACM Transactions on Graphics, vol. 35, no. 6, pp. 193:1–193:10, Nov 2016.
[56] B. E. Bayer, "Color Imaging Array," U.S. Patent 3 971 065, Jul 20, 1976.
[57] K. Hirakawa and P. J. Wolfe, "Spatio-Spectral Color Filter Array Design for Optimal Image Recovery," IEEE Transactions on Image Processing, vol. 17, no. 10, pp. 1876–1890, Oct 2008.
[58] L. Condat, "A New Color Filter Array With Optimal Properties for Noiseless and Noisy Color Image Acquisition," IEEE Transactions on Image Processing, vol. 20, no. 8, pp. 2200–2210, Aug 2011.
[59] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," Foundations and Trends in Machine Learning, vol. 3, no. 1, pp. 1–122, 2011.
[60] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A Fast Approach for Overcomplete Sparse Decomposition Based on Smoothed ℓ^0 Norm," IEEE Transactions on Signal Processing, vol. 57, no. 1, pp. 289–301, Jan 2009.
[61] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcompletedictionaries for sparse representation," IEEE Transactions on Signal Processing, vol. 54, no. 11, Nov. 2006.
[62] E. Miandji, "Sparse representation of visual data for compression and compressed sensing," Ph.D. dissertation, Linköping University, Media and Information Technology, Faculty of Science & Engineering, 2018.
[63] D. L. Snyder, A. M. Hammoud, and R. L. White, "Image Recovery from Data Acquired with a Charge-coupled-device Camera," Journal of the Optical Society of America A, vol. 10, no. 5, pp. 1014–1023, May 1993.
[64] H. Lantéri and C. Theys, "Restoration of astrophysical images – the case of poisson data with additive gaussian noise," EURASIP Journal on Advances in Signal Processing, p. 643143, Sep 2005.
[65] F. Benvenuto, A. L. Camera, C. Theys, A. Ferrari, H. Lantéri, and M. Bertero, "The Study of an Iterative Method for The Reconstruction of Images Corrupted by Poisson and Gaussian noise," Inverse Problems, vol. 24, no. 3, p. 035016, Apr 2008.