Light Field Synthesis by Training Deep Network in the Refocused Image Domain

Chang-Le Liu, Kuang-Tsu Shih, Jiu-Woei Huang, and Homer H. Chen, Fellow, IEEE

Abstract—Light field imaging, which captures spatio-angular information of light incident on image sensor, enables many interesting applications such as image refocusing and augmented reality. However, due to the limited sensor resolution, a trade-off exists between the spatial and angular resolutions. To increase the angular resolution, view synthesis techniques have been adopted to generate new views from existing views. However, traditional learning-based view synthesis mainly considers the image quality of each view of the light field and neglects the quality of the refocused images. In this paper, we propose a new loss function called refocused image error (RIE) to address the issue. The main idea is that the image quality of the synthesized light field should be optimized in the refocused image domain because it is where the light field is viewed. We analyze the behavior of RIE in the spectral domain and test the performance of our approach against previous approaches on both real and software-rendered light field datasets using objective assessment metrics such as MSE, MAE, PSNR, SSIM, and GMSD. Experimental results show that the light field generated by our method results in better refocused images than previous methods.

Index Terms—light field, view synthesis, CNN, image refocusing

I. INTRODUCTION

Light field imaging enables users to refocus and change view-angle. A light field can be captured by a light field camera equipped with a microlens array [1], [2] or by a camera array [3]. The angular resolution of the light field captured by the former is limited by the number of pixels covered by microlens and the latter by the number of cameras. In practice, only limited angular resolution is available. To increase the angular resolution, view synthesis (interpolation) is often adopted to generate new views from the existing light field.

A view synthesis approach that has been commonly adopted first estimates the disparity map (or depth map) of the light field and then uses it to generate new views in between the existing views by warping the existing light field [4]–[7]. Recently, deep learning has been applied to view synthesis of light field [8]–[10]. It involves several convolutional neural networks (CNN) to estimate the disparity map and refine the warped new views using an end-to-end training strategy. The output is a light field denser than the input. The loss functions used are mostly $L1$ or $L2$ metric. Furthermore, these metrics for evaluating the quality of the synthesized light field is performed view-wise (that is, view by view) in the 4D light field domain.

In our view, the quality of view synthesis for light field should be evaluated in the refocused image domain, because what matters ultimately is the perceived image, not the volume of 4D raw data. Since the perceived image is a refocused image, we believe a light field synthesis optimized in the refocused domain would generate better refocused images. Fig. 1 illustrates a common drawback of conventional light field synthesis. View-wise optimization usually results in artifact at the region where occlusion occurs.

To take the refocused image quality into consideration, we propose to add refocused image error (RIE) to the traditional view-wise loss function as a regularization term. RIE encourages the network to focus on the light field quality in the refocused image domain. It results in a dense light field, from which a high-quality refocused image can be generated.

The contributions of this paper can be summarized as follows:

- To our best knowledge, this is the first work that considers refocused image quality for light field synthesis and optimizes the deep network in the refocused image domain.
- We analyze the proposed refocused image error in the spectral domain and show the relation between evaluating $L2$ loss in

This work was supported in part by a grant from the Ministry of Science and Technology of Taiwan under Contract 106-2221-E-002-201-MY3, and in part by grants from National Taiwan University under Contracts CC-NTU: 108L891808 and 108L880502.

C.-L. Liu is with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: b05901017@ntu.edu.tw)
K.-T. Shih is with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: shihkt@gmail.com)
J.-W. Huang is with the Institute of Applied Mechanics, National Taiwan University, Taipei 10617, Taiwan (jwhuang@ms2.hinet.net)
H. H. Chen is with the Department of Electrical Engineering, Graduate Institute of Communication Engineering, and Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan (e-mail: homer@ntu.edu.tw).

Fig. 1. Illustration of artifacts generated by a network trained to minimize individual view loss. (a) A new view synthesized from a light field in the HCI dataset. (b) The blowup view of a small region.
4D light field domain and that in the refocused image domain. 
* We demonstrate that taking refocused image quality into consideration improves the performance of deep learning-based light field synthesis.

The rest of the paper is organized as follows. We review the related work in Sec. II and introduce the notation used in this paper in Sec. III. The proposed regularization and related analysis are described in Sec. IV. The experiment setting and the results are described in Sec. V. Finally, the concluding remarks and future work discussion are provided in Sec. VI.

II. RELATED WORK

A. Light Field Imaging

A light field [11] records the spatio-angular information of the light rays coming from different view angles. It contains angular information unavailable in the traditional 2D image data. Using the spatio-angular information, we may perform image refocusing to make any object in the scene in focus [1]. A light field can be captured by using an array of cameras [3] or a camera with a lenslet array [1]. The latter allows miniaturization of the device for consumer electronics [12], [13]. However, both types of light field cameras have limited angular resolution. The angular resolution of the former is limited by the number of cameras and the latter by the number of corresponding pixels of a lenslet.

B. View Synthesis for Light Field

View synthesis [3] has been developed to increase the angular resolution of a light field. Generally, view synthesis methods for the light field could be classified into two types.

The first type of methods [4], [5], [7] first estimate the depth information and then warp the existing views to generate new views by multi-view stereo algorithms [14], [15]. It is a depth-dependent process. The convolutional neural network (CNN) has also been adopted for view synthesis. Kalantari et al. proposed to use CNN to evaluate disparity information from the four corner views in an input light field. Another CNN was used to refine the new views generated the existing views and to estimate the depth map [9]. Srinivasan et al. extended this method to synthesize a light field only from the central view in the light field using an extra constraint on the consistency of ray depths [8].

The second type of methods synthesize a new image without depth information. This is made possible by limiting the configuration of input views to some specific patterns [16], [17]. Although the depth information is not required, the constraint on the configuration of input views limits the application of such methods. To overcome this problem and improve the performance for occlusion regions and non-Lambertian surfaces (which do not reflect light equally in all directions), for which depth-dependent methods often fail, Wang et al. proposed Pseudo 4DCNN [10] that adopts 3D CNN [18] to extract the 3D volume features by alternatively fixing an angular dimension of the input light field. That is, the Pseudo 4DCNN directly generate a dense light field by upsampling the input light field using deconvolution.

Although the above methods solve the light field view synthesis problem to a certain extent, the quality of refocused images is not explored. In this paper, we propose to add a regularization term called refocused image error to the loss function. It encourages the network to generate high-quality refocused images.

III. NOTATION

We consider the 4D light field proposed by Levoy and Hanrahan [3]. Denote a light field by \( L(x, y, s, t) \), where the pairs \((x, y)\) and \((s, t)\) represent spatial and angular coordinates respectively. In practice, both \( s \) and \( t \) are finite; therefore, we assume each of them is bounded by \( N \). Because a 4D light field can be considered an image array captured from different viewpoints, we use \( L_s(x, y) \), where \( s = (x, y) \), to denote the view captured at the angular coordinates \( s = (s, t) \). That is, \( L_s(x) \) is a sub-aperture image at \((s, t)\).

The shift-and-add is the basic operation of image refocusing. In this operation, we first shift each sub-aperture image \( L_s(x) \) by \( \Delta x = rs \) and then average all shifted image to generate a refocused image \( R \). More specifically,

\[
R(L, r, x) = \frac{1}{(2N+1)^2} \sum_{s} L_s(x + rs).
\]

Note that a larger \( r \) means refocusing farther from the camera and the sign of \( r \) decides the position of the refocused focal point with respect to the original focal point: positive means refocusing at a farther object and negative at a closer object.

In Sec. IV, we use \( \mathfrak{S} \) and \( \mathfrak{S}^{-1} \) to denote non-unitary Fourier transform and inverse Fourier transform. Note that we drop the constant coefficients like \( 1/2\pi \) in \( \mathfrak{S}^{-1} \) for simplicity, and this simplification does not affect our analysis. We use \( \text{sinc}(x) \) to denote the unnormalized sinc function \( \text{sinc}(x) = \sin(x)/x \).

IV. REFOCUSED IMAGE ERROR

In this section, we first introduce the refocused image error and provide related analysis in the frequency domain.

A. Refocused Image Error

Assume there exists a desired light field \( L \) and a set of images \( S \) sampled from \( L \). Given input \( S \), we want to train a neural network \( G \) parameterized by \( \theta \) to predict a light field \( G_\theta(S) \) that is as similar as possible to \( L \). Note that the proposed method has no restriction on the type of input, even though the input of the network is a set of views of the desired light field in this work. The only limitation is that the output must be a light field. Mathematically, the network \( G \) is trained to minimize the loss between \( G_\theta(S) \) and \( L \) as follows:

\[
\theta = \arg \min_\theta \mathcal{L}(G_\theta(S), L)
\]

Traditionally, the loss function \( \mathcal{L} \) is chosen to be the mean-squared error (MSE) or the mean absolute error (MAE) between every image in \( L \) and \( G_\theta(S) \). These view-wise error (VWE) can be defined as follows:

\[
\text{VWE}_s(G_\theta(S), L) = \frac{1}{N^2} \sum_{s} \text{MAE}(G_\theta(S), L_s)
\]
This kind of loss functions only encourage the network to perform well on each sub-aperture image without considering the quality of refocused images generated from the predicted light field $G_b(S)$. Instead, we propose the unweighted continuous refocused image error (UCRIE),

$$UCRIE_1(G_b(S), L) = \frac{1}{2D} \int_{-D}^{D} g(r) \text{MAE}(R(G_b(S), r), R(L, r))dr,$$

$$UCRIE_2(G_b(S), L) = \frac{1}{2D} \int_{-D}^{D} g(r) \text{MSE}(R(G_b(S), r), R(L, r))dr,$$

where $D$ denotes the maximal value of $r$ in the shift-and-add operation. Intuitively, UCRIE$_2$ and UCRIE$_1$ correspond to MSE and MAE, respectively. Also, it has been found empirically that the quality of a refocused image in (1) with $r$ close to zero is more important for learning; therefore, we weight UCRIE by a Gaussian function $g(r) = \exp(-r^2)$ and define the continuous refocused image error (CRIE) as follows:

$$CRIE_1(G_b(S), L) = \frac{1}{2D} \int_{-D}^{D} g(r) \text{MAE}(R(G_b(S), r), R(L, r))dr,$$

$$CRIE_2(G_b(S), L) = \frac{1}{2D} \int_{-D}^{D} g(r) \text{MSE}(R(G_b(S), r), R(L, r))dr,$$

Note that the definite integral is replaced with improper integral in CRIE for simplification in the spectral domain. (By the definition of $g(r)$, this makes a trivial difference when $D$ is larger than 3.) Incidentally, we adopt definite integral for UCRIE to prevent it from diverges when $D$ goes to infinity. An analysis of CRIE in the spectral domain is given in Sec. IV.B to explain why we add $g(r)$.

Although $r$ in UCRIE and CRIE can have infinitely many values, evaluating an equation on infinitely many points is difficult. In practice, the following losses, called refocused image errors (RIEs) in this work, are more appropriate for deep learning and quality evaluation tasks:

$$RIE_1(G_b(S), L) = \frac{1}{2D} \sum_{s=-D}^{D} g(r) \text{MAE}(R(G_b(S), sr), R(L, sr)),$$

$$RIE_2(G_b(S), L) = \frac{1}{2D} \sum_{s=-D}^{D} g(r) \text{MSE}(R(G_b(S), sr), R(L, sr)),$$

where $s$ is the step interval of the summation.

Fig. 2. (a) and (b) Different sub-aperture images chosen from a 5x5 light field. (c) and (d) The superposition of two sub-aperture images in (a) and (b). (e) A 135-degree orientated Gaussian function. (f) A 60-degree orientated Gaussian function.

B. Spectral Domain Analysis

Here we analyze the characteristics of CRIE in the spectral domain. First, we may rewrite CRIE$_2$ using Fourier transform and Plancherel’s formula as follows:

$$UCRIE_2(G_b(S), L) = \frac{1}{(2N+1)^2} \sum_{x} \sum_{\omega} \mathcal{F}_x(\omega) \mathcal{F}_x^*(\omega) \text{sinc}(D\omega^T (s+t)),$$

where $\mathcal{F}_x = \mathfrak{F}\{G_b(S) - L_x\}$ denotes the spectrum of the error of the sub-aperture image at $s$. Similarly, CRIE$_2$ can be rewritten as

$$CRIE_2(G_b(S), L) = \frac{1}{(2N+1)^2} \sum_{x} \sum_{\omega} \mathcal{F}_x(\omega) \mathcal{F}_x^*(\omega) \frac{\sqrt{\pi}}{2D} e^{-0.25(\omega^T (s+t))^2}$$

The derivation of Eqs. (11) and (12) is given in Appendix.
Eqs. (11) and (12) suggest that the refocused continuous refocused image error measures the error filtered by a low-pass filter: sinc filter for CRIE_2 and Gaussian filter for CRIE_2. By the definition of MSE, we can also rewrite (3) in frequency domain as

\[
\sum_{s \in \mathbb{N}^2} \text{MSE}(L_s, G_p(S)_s) = \sum_{s \in \mathbb{N}^2} \sum_{\omega} \mathcal{E}_s(\omega)\mathcal{E}_t(\omega). \tag{13}
\]

We can see that view-wise MSE is a simplification of Eq. (12). Compare both CRIE_2 and UCRIE_2 with traditional view-wise MSE, we can see there are two main differences. First the traditional view-wise MSE does not consider the correlation between two different views captured from different view angles. Second, the errors measured by CRIE_2 and UCRIE_2 are filtered by a directional filter. That is, the weighting for each \(\sum_{s \in \mathbb{N}^2} \mathcal{E}_s(\omega)\mathcal{E}_t(\omega)\) in Eqs. (11) and (12) depends on the vector value of \(s + t\). Take CRIE_2 for example and let \(s = (0, 0)\) W.L.O.G. Assume \(s + t = (1, 1)\). Then \(L_s\) is in the top right of \(L_s\), as marked in the Fig. 2(a). According to Eq. (12), we see the weight of \(\mathcal{E}_s(\omega)\mathcal{E}_t(\omega)\) as an 135-degree orientated low-pass Gaussian filter, as shown in Fig. 2(e). This is reasonable. If we shift-and-add \(L_s\) and \(L_t\); \(L_s\) moves toward (away) from \(L_s\) when \(r\) is positive (negative) in the 45-degree orientation, we get an image with motion blur in the 45-degree orientation, as shown in Fig. 2(c). In other case, assume \(s + t = (2, -1)\), as shown in Fig. 2(f). Then the weight is a 60-degree orientated low-pass Gaussian filter Fig. 2(f), and the superposition of \(L_s\) and \(L_t\), as shown in Fig. 2(d).

The above analysis also explains why CRIE_2 is better than UCRIE_2. The Gaussian filter is non-negative and non-oscillatory and hence causes no ringing effect in the spectral domain.

For CRIE_3 and UCRIE_3, we can analyze them in a similar way. By using Chebyshev approximation [19] and omitting terms higher than the fourth order, we have the following:

\[
\text{CRIE}_1(G_p(S), L) = \frac{-1}{\pi} \frac{1}{2D(2N+1)} \int_{-\infty}^{\infty} e^{-r^2} \left\{ \sum_{k} \left( \frac{1}{\sqrt{2\pi}} \frac{1}{2} T_{2k}(\tau(x,r)) \right) e^{ix} \right\} \text{d}r,
\]

where \(T_{2k}\) is a Chebyshev polynomial of the first kind. By expanding (14), we can find that CRIE_3 includes CRIE_2.

V. EXPERIMENTAL SETUP AND RESULTS

In this section, we first describe the network architecture used in our experiment for view synthesis and the experimental setup. Then, we describe the results of our experiments for testing the performance of the proposed RIE for light field synthesis.

A. Network Architecture

We consider a neural network trained through backpropagation for light field synthesis. Specifically, the deep neural network architecture proposed by Srinivasan et al. [8] was used in the experiments. Since synthesizing a light field from multiple views is more robust than from a single view, we fed a total of five views into the network as inputs. This architecture consists of two sub-networks, both fully convolutional networks. The first sub-network estimated the depth map of each new view, based on which an approximate Lambertian light field was synthesized by warping the central view. The second sub-network predicted a residual light field to be added to the synthesized Lambertian light field and handled occluded parts and non-Lambertian effects [8]. In our experiment, we trained two networks to minimize the following two loss functions, one for each network:

\[
\text{VWE}_1(G_p(S), L) + \lambda_R \text{RIE}_1(G_p(S), L), \tag{15}
\]
\[
\text{VWE}_2(G_p(S), L) + \lambda_R \text{RIE}_2(G_p(S), L), \tag{16}
\]

where \(\lambda_R\) denotes the RIE regularization parameters. The training scheme is shown in Fig. 3. For comparison, we trained another two networks to minimize Eqs. (3) and (4).
Note that at first glance, using Eqs. (11) and (12) as the loss function may seem appropriate for evaluating Eqs. (7) and (8) on infinitely many samples of \( r \). However, using them for training usually leads to unstable results. This is why Eqs. (9) and (10) are used in Eqs. (15) and (16), respectively, for network training in our experiments.

B. Experimental Setup

We evaluated the network performance in terms of following standard objective metrics: MAE, MSE, the peak signal-to-noise ratio (PSNR), the gray-scale structural similarity (SSIM) [20], and the gradient magnitude similarity deviation (GMSD) [21]. SSIM is one of perception-based image quality metric that considers image luminance, contrast, and structure similarity. The definition of SSIM is described below:

\[
\text{SSIM}(x, y) = [(l(x, y))^{2}][c(x, y)]^{2}[s(x, y)]^{2}
\]
\[
l(x, y) = \frac{2\mu_{x}\mu_{y} + C_{1}}{\mu_{x}^{2} + \mu_{y}^{2} + C_{1}}
\]
\[
c(x, y) = \frac{2\sigma_{x}\sigma_{y} + C_{2}}{\sigma_{x}^{2} + \sigma_{y}^{2} + C_{2}}
\]
\[
s(x, y) = \frac{\sigma_{xy} + C_{3}}{\sigma_{x}\sigma_{y} + C_{3}}
\]

(17)

where \( \mu_{x}, \mu_{y}, \sigma_{x}, \sigma_{y}, \) and \( \sigma_{xy} \) are the local means, standard deviations, and cross-covariance for distorted images \( x \) and the reference image \( y \). \( C_{1}, C_{2}, \) and \( C_{3} \) are variables to stabilize the division. GMSD is another image quality metric that uses the standard deviation of the pixel-wise gradient magnitude similarity (GMS) to evaluate image quality,

\[
\text{GMS}(x) = \frac{2m_{x}(x)m_{y}(x) + c}{m_{x}(x) + m_{y}(x) + c},
\]

\[
\text{GMSD} = \frac{1}{M} \sum_{x} \text{GMS}(x) - \frac{1}{M} \sum_{x} \text{GMS}(x)^{2},
\]

(18)

where \( M \) is the number of pixels, \( m_{x} \) and \( m_{y} \) are gradient magnitudes of the reference and the distorted images, respectively. The worse the distorted image, the higher GMSD is. We used SSIM and GMSD because these image quality metrics have high agreement with the subjective experimental results [22] for light field quality evaluation.

We trained the networks on two light field datasets: the synthetic light fields (HCI dataset) [23] and the real light fields (INRIA dataset [24]). We partitioned the HCI dataset into sixteen light fields for training and eight for testing. Likewise, we partitioned the INRIA dataset into 43 light fields for training and 16 for testing. Both HCI and INRIA datasets have a spatial resolution of 512 \( \times \) 512. But the former has angular resolution of 9 \( \times \) 9, and the latter 7 \( \times \) 7. Furthermore, we extracted sub-lightfields of 5 \( \times \) 5 views from each dataset. A total of 25 sub-lightfields were extracted from the HCI dataset and nine from the INRIA dataset. For each sub-light field, we used the central view and the four corner views as input to the view synthesis network. The output was a 3 \( \times \) 3 light field, as shown in Fig. 3.

| TABLE I | 4D LIGHT FIELD QUALITY COMPARISON OF OUR APPROACH WITH TRADITIONAL LOSS FUNCTION USING HCI DATASET |
|---|---|---|---|---|
| Quality Metrics | VWE | VWE + RIE | VWE | VWE + RIE |
| MAE | 0.0180 | 0.0176 | 0.0152 | 0.0156 |
| MSE | 0.0032 | 0.0029 | 0.0012 | 0.0010 |
| PSNR | 27.310 | 28.098 | 30.340 | 30.727 |

| TABLE II | 4D LIGHT FIELD QUALITY COMPARISON OF OUR APPROACH WITH TRADITIONAL LOSS FUNCTION USING INRIA DATASET |
|---|---|---|---|---|
| Quality Metrics | VWE | VWE + RIE | VWE | VWE + RIE |
| MAE | 0.0176 | 0.0094 | 0.0091 | 0.0091 |
| MSE | 0.0030 | 0.0003 | 0.0005 | 0.0004 |
| PSNR | 27.805 | 36.397 | 33.354 | 35.213 |

| TABLE III | REFOCUSED IMAGE QUALITY COMPARISON OF OUR APPROACH WITH TRADITIONAL LOSS FUNCTION USING HCI DATASET |
|---|---|---|---|---|
| Quality Metrics | VWE | VWE + RIE | VWE | VWE + RIE |
| MAE | 0.0096 | 0.0091 | 0.0071 | 0.0070 |
| MSE | 0.0008 | 0.0007 | 0.0002 | 0.0002 |
| PSNR | 31.087 | 31.283 | 37.183 | 38.308 |
| SSIM | 0.8243 | 0.8358 | 0.8468 | 0.8415 |
| GMSD | 0.0461 | 0.0371 | 0.0269 | 0.0240 |

| TABLE IV | REFOCUSED IMAGE QUALITY COMPARISON OF OUR APPROACH WITH TRADITIONAL LOSS FUNCTION USING INRIA DATASET |
|---|---|---|---|---|
| Quality Assessment | VWE | VWE + RIE | VWE | VWE + RIE |
| Used loss function in training | MAE | VWE | VWE + RIE | VWE | VWE + RIE |
| MAE | 0.0093 | 0.0043 | 0.0043 | 0.0041 |
| MSE | 7.08e-4 | 4.00e-5 | 6.82e-5 | 4.12e-5 |
| PSNR | 31.523 | 44.0113 | 41.7215 | 43.8753 |
| SSIM | 0.8271 | 0.9129 | 0.9143 | 0.9163 |
| GMSD | 0.0428 | 0.0055 | 0.0132 | 0.0074 |

For each new view, the corresponding view at the same position in the original light field was used as the ground-truth. For the RIE evaluation, we set \( D = 2.5, s = 0.25 \) and \( \lambda_r = 1 \). All neural network models were trained using the Adam optimization algorithm [25] with default parameter values \( \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e^{-8} \).

C. Experimental Results

We trained two networks using the proposed loss functions. The one using Eq. (15) as the loss function is referred to as \( \text{VWE}_1 + \text{RIE}_1 \), and the one using Eq. (16) as the loss function is referred to as \( \text{VWE}_2 + \text{RIE}_2 \). In addition, two other networks called \( \text{VWE}_1 \) and \( \text{VWE}_2 \) for short were created using Eqs. (3) and (4), respectively, as the loss functions. We also trained
another network using the state-of-the-art loss function proposed by Srinivasan et al. [8]. The latter three networks served as the baseline for comparison, and all five networks were tested on the HCI and INRIA datasets. We measured the mean scores of all sub-aperture views using MAE, MSE and PSNR as the quality metrics and summarized the results in Tables I and II. We can see the PSNR and MSE scores of VWE₂ + RIE₂ and VWE₁ + RIE₁, respectively, are higher than those of VWE₂ and VWE₁. The MAE score of VWE₂ + RIE₂ is also higher than that of VWE₂. In addition, both VWE₁ + RIE₁ and VWE₁ perform better than VWE₂ + RIE₂ and VWE₂. Overall, MAE is a better loss function than MSE for light field synthesis. The results are in line with those reported by Zhao et. al. [26] for image restoration.

Blowup views of the outputs of the trained networks and the corresponding ground-truth are shown in Fig. 4 for the HCI dataset and Fig. 5 for the INRIA dataset. We can see that the results of VWE₁ + RIE₁ is less noisy and more stable than those of VWE₁ and VWE₂. From Fig. 4(b), we note that the artifact of VWE₁ and VWE₂ in the occlusion region is quite pronounced, while the networks VWE₁ + RIE₁ with the proposed loss function is almost artifact-free. From Fig. 5, we can see that VWE₁ and VWE₂ generated more noises than VWE₁ + RIE₁.

In addition to sub-aperture view quality, we evaluated the refocused image quality. The MAE, MSE, PSNR, SSIM, and GMSD scores of the refocused images generated by the five networks with r = 0 are shown in Tables III and IV. We can see from Table III that VWE₁ + RIE₁ and VWE₁ + RIE₂ outperform VWE₂ and VWE₁ in terms of MAE, PSNR, and GMSD when
evaluated on the HCI dataset. Furthermore, the performance of $VWE_2+RIE_2$ and $VWE_1+RIE_1$ is superior to the other three networks by all five metrics when tested on the INRIA dataset. It is worth noting that the average MAE, MSE, and PSNR of the refocused image are lower than the sub-aperture views. This can be explained by Eq (1) and the triangle inequality. When we obtain a refocused image from a light field, the errors of different sub-aperture views may cancel each other due to the shift-and-sum operation.

The quality of refocused images generated from the synthesized light fields with $r$ ranging from $-2.5$ to $2.5$ are also evaluated. The results summarized in Fig. 7 clearly show that $VWE_2+RIE_2$ and $VWE_1+RIE_1$ perform better than $VWE_2$ and $VWE_1$ by all five image metrics. This is reasonable because a refocused image of a light field can be seen as a linear combination of the shifted sub-aperture images. Therefore, when the sub-aperture images are noisy, the refocused images are likely to be noisy as well. From Figs. 4−7 and Tables I−IV, we can see that the networks trained by the proposed $VWE_1+RIE_1$ functions indeed give rise to superior light field synthesis in both the 4-D light field domain and the refocused image domain.

VI. CONCLUSION

In this paper, we have described a novel loss called refocused image error for light field synthesis. It drives a deep network to minimize light field loss in the 4D light field domain and the refocused image domain at the same time, resulting in high-quality refocused images. The superior performance of the proposed loss is supported by a theoretical analysis that shows the refocused image error is related to the summation of the inner products of spectra errors between all view pairs of a synthesized light field. In effect, our technique performs a global optimization.

The experimental results using real and software-rendered light field datasets clearly show that the proposed regularization is more effective than the conventional one that only considers the individual view quality of a light field. The proposed loss is potentially useful for other light-field related tasks such as light field compression [27] and super-resolution [28]. These topics are worth further investigation in the future.
APPENDIX

For simplicity, let \( \hat{L} \) denote the alias of \( G(S) \) and \( \hat{L}^* = L_x(x + h) \). The definitions of UCRIE\(_2\) and shift-and-add operator in Eqs. (6) and (1) establish the equation:

\[
\text{UCRIE}_2(\hat{L}, L) = \frac{1}{2D} \int_D \int_D \text{MSE}(R(\hat{L}, r), R(L, r))dr
\]

\[
= \frac{1}{2D} \int_D \sum_x (R(\hat{L}, r, x) - R(L, r, x))^2dr
\]

\[
= \frac{1}{2D} \int_D \sum_x (\frac{1}{(2N+1)^2} \sum_{s} \hat{L}_s^*(x + rs) - \frac{1}{(2N+1)^2} \sum_{s} L_s^*(x + rs))^2dr
\]

\[
= \frac{1}{2D(2N+1)^2} \int_D \sum_x \sum_{s} (\hat{L}_s^*(x) - L_s^*(x))^2dr
\]

Fig. 6. Refocused images and close-ups from the ground-truth and the predicted light fields.
Because the light fields are finite-valued, we can interchange the order of summation:

\[
\frac{1}{2D(2N+1)^4} \int_{D} \sum_{x} \left( \sum_{y} \left( \hat{L}_{x}^{y}(x) - L_{x}^{y}(x) \right)^2 \right) \, dr
\]

\[
= \frac{1}{2D(2N+1)^4} \int_{D} \sum_{y} \sum_{x} \left( \hat{L}_{x}^{y}(x) - L_{x}^{y}(x) \right) \left( \hat{L}_{x}^{y}(x) - L_{x}^{y}(x) \right) \, dr
\]
Let $\mathcal{E}_s = \mathcal{F}\{\hat{L}_s - L_s\}$ for simplicity. By Plancherel’s formula and the translation property of Fourier transform, we have

$$\frac{1}{2D(2N+1)^2} \int_{-D}^D \sum_{k,t} \left(\hat{L}_s(x) - L_s(x)\right) \left(\hat{L}_t(x) - L_t(x)\right) dx$$

$$= \frac{1}{2D(2N+1)^2} \int_{-D}^D \sum_{s,t} \left(\mathcal{E}_s(\omega)e^{j\omega s \cdot x}\right) \left(\mathcal{E}_t(\omega)e^{j\omega t \cdot x}\right) dx$$

The final step is to interchange the summation and the integration again,

$$\frac{1}{2D(2N+1)^2} \sum_{s,t} \sum_{\omega} \mathcal{E}_s(\omega) \mathcal{E}_t(\omega) \frac{1}{2D} \int_{-D}^D e^{j\omega (s + t) \cdot x} dx$$

$$= \frac{1}{(2N+1)^2} \sum_{s,t} \mathcal{E}_s(\omega) \mathcal{E}_t(\omega) \frac{\pi}{2D} e^{-0.25\omega (s + t) \cdot 2}.$$

This completes the proof. 

□

References

[1] R. Ng et al., “Light Field Photography with a Hand-held Plenoptic Camera,” p. 11.

[2] T. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. Nayar, and C. Intwala, “Spatio-Angular Resolution Tradeoff in Integral Photography,” p. 10.

[3] M. Levoy and P. Hanrahan, “Light field rendering,” in Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH ’96, Not Known, 1996, pp. 31–42.

[4] G. Chaurasia, O. Sorkine, and G. Drettakis, “Silhouette-Aware Warping for Image-Based Rendering,” Comput. Graph. Forum, vol. 30, no. 4, pp. 1223–1232, Jun. 2011.

[5] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis, “Depth synthesis and local warps for plausible image-based navigation,” ACM Trans. Graph., vol. 32, no. 3, pp. 1–12, Jun. 2013.

[6] S. Wanner and B. Goldluecke, “Variational Light Field Analysis for Disparity Estimation and Super-Resolution,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 3, pp. 606–619, Mar. 2014.

[7] M. Goesele, J. Ackerman, S. Fuhrmann, C. Haubold, R. Klowsky, and T. Darmstadt, “Ambient point clouds for view interpolation,” ACM Trans. Graph., vol. 29, no. 4, p. 1, Jul. 2010.

[8] P. P. Srinivasan, T. Wang, A. Sreeval, R. Ramamoorthi, and R. Ng, “Learning to Synthesize a 4D RGBD Light Field from a Single Image,” ArXiv170803292 Cs, Aug. 2017.

[9] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, “Learning-based view synthesis for light field cameras,” ACM Trans. Graph., vol. 35, no. 6, pp. 1–10, Nov. 2016.

[10] Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, and T. Tan, “End-to-End View Synthesis for Light Field Imaging with Pseudo 4DCNN,” in Computer Vision – ECCV 2018, vol. 11206, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 340–355.

[11] G. Lippmann, “La photographie integrale,” ComptesRendus Acad. Sci., 1908.

[12] “Lytro.” [Online]. Available: https://www.lytro.com/.

[13] “Raytrix,” 2018. [Online]. Available: https://raytrix.de/.

[14] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz, “Multi-View Stereo for Community Photo Collections,” in 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[15] Y. Furukawa and J. Ponce, “Accurate, Dense, and Robust Multi-View Stereopsis,” p. 8.

[16] A. Levin and F. Durand, “Linear view synthesis using a dimensionality gap light field prior,” in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 1831–1838.

[17] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, “Light Field Reconstruction Using Sparsity in the Continuous Fourier Domain,” ACM Trans. Graph., vol. 34, no. 1, pp. 1–13, Dec. 2014.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks,” in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 4489–4497.

[19] P. R. Clement, “The Chebyshev approximation method,” Q. Appl. Math., vol. 11, no. 2, pp. 167–183, Jul. 1953.

[20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image Quality Assessment: From Error Visibility to Structural Similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[21] W. Xue and L. Zhang, “Gradient Magnitude Similarity Deviation: An Highly Efficient Perceptual Image Quality Index,” p. 12.

[22] V. K. Adhikarla et al., “Towards a Quality Metric for Dense Light Fields,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 3720–3729.

[23] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, “A Dataset and Evaluation Methodology for Depth Estimation on 4D Light Fields,” in Computer Vision – ACCV 2016, vol. 10113, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham: Springer International Publishing, 2017, pp. 19–34.

[24] X. Jiang, M. Le Pendu, R. A. Farrugia, and C. Guillemot, “Light Field Compression With Homography-Based Low-
Chang-Le Liu is currently working toward the B.S. degree in electrical engineering in National Taiwan University, Taipei, Taiwan, from 2016. He was a summer intern as a research assistant with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan, and was involved in research in speech enhancement. His current research topic includes image and audio signal processing and light field photography and rendering.

Kuang-Tsu Shih received the Ph.D. degree in communication engineering from National Taiwan University, Taipei, Taiwan, in 2017. He is currently a Postdoctoral Researcher with the Graduate Institute of Communication Engineering, National Taiwan University. His current research interests include color image processing, color science, and computational photography.

Jiun-Woei Huang is Research Fellow of Taiwan Instrument Research Institute, National Applied Laboratories, Taiwan, R.O.C., and adjunct associate professor, Institute of Applied Mechanics, National Taiwan University, since 2008. He received his BS and MS degrees in physics from the National Taiwan Normal University in 1977 and 1981, respectively, and his Ph.D. degree in Physics from Texas Christian University, USA, with Professor W.R.M. Graham in Molecular Spectroscopy, in May 1990. He worked as Post Doc of Lidar Laboratory of Professor C. Y. She in Colorado State University from May, 1990 to Aug. 1991, then back to Taiwan. From late 1991 to Jan. 1992. He was a Post-Doc Researcher in the Institute of Chemistry of Academia Sinica, R.O.C., studied for Electron Paramagnetic Resonance in minerals. Then he had been in CSIST from 1992 to 2014 in the field of optical design, mechanical structure, fiber gyro, and infrared optics. He is a member of SPIE, OSA, and life member of Sigma Xi.

Homer H. Chen (S’83–M’86–SM’01–F’03) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign. His professional career has spanned industry and academia. Since August 2003, he has been with the College of Electrical Engineering and Computer Science, National Taiwan University, where he is a Distinguished Professor. Prior to that, he held various research and development management and engineering positions at U.S. companies over a period of 17 years, including AT&T Bell Labs, Rockwell Science Center, iVast, and Digital Island. He was a U.S. delegate for ISO and ITU standards committees and contributed to the development of many new interactive multimedia technologies that are now a part of the MPEG-4 and JPEG-2000 standards. His professional interests lie in broad areas of multimedia signal processing and communications.

Dr. Chen was a Distinguished Lecturer of the IEEE Circuits and Systems Society from 2012 to 2013. He served on the IEEE Signal Processing Society Fourier Award Committee and the Fellow Reference Committee from 2015 to 2017. He serves on the IEEE Signal Processing Society Awards Board and the Senior Editorial Board of the IEEE Journal on Selected Topics In Signal Processing. He is a General Chair of the 2019 IEEE International Conference on Image Processing. He was an Associate Editor of the IEEE Transactions on Circuits and Systems For Video Technology from 2004 to 2010, the IEEE Transactions on Image Processing from 1992 to 1994, and Pattern Recognition from 1989 to 1999. He served as a Guest Editor for the IEEE Transactions on Circuits and Systems For Video Technology in 1999, the IEEE Transactions on Multimedia in 2011, the IEEE Journal of Selected Topics In Signal Processing in 2014, and Multimedia Tools and Applications (Springer) in 2015.