Self-Supervised Light Field Reconstruction Using Shearlet Transform and Cycle Consistency

Yuan Gao, Robert Bregović, *Member, IEEE*, and Atanas Gotchev, *Member, IEEE*

***Abstract*—Shearlet Transform (ST) has been instrumental for the Densely-Sampled Light Field (DSLF) reconstruction, as it sparsifies the underlying Epipolar-Plane Images (EPIs). The sought sparsification is implemented through an iterative regularization, which tends to be slow because of the time spent on domain transformations for dozens of iterations. To overcome this limitation, this letter proposes a novel self-supervised DSLF reconstruction method, CycleST, which employs ST and cycle consistency. Specifically, CycleST is composed of an encoder-decoder network and a residual learning strategy that restore the shearlet coefficients of densely-sampled EPIs using EPI-reconstruction and cycle-consistency losses. CycleST is a self-supervised approach that can be trained solely on Sparsely-Sampled Light Fields (SSLFs) with small disparity ranges ($\leqslant$ 8 pixels). Experimental results of DSLF reconstruction on SSLFs with large disparity ranges (16 - 32 pixels) demonstrate the effectiveness and efficiency of the proposed CycleST method. Furthermore, CycleST achieves $\sim$9x speedup over ST, at least.**

***Index Terms*—Image-based rendering, light field reconstruction, self-supervision, shearlet transform, cycle consistency.**

I. Introduction

DENSELY-Sampled Light Field (DSLF) is a discrete 4D representation [1, 2] for the light rays from the scene encoded by two parallel planes, namely image plane and camera plane, where the maximum disparity between neighboring views is one pixel at most [3, 4]. DSLF has a wide range of applications, *e.g.*, synthetic aperture imaging, depth estimation and visual odometry [5]. DSLF-based contents can be rendered on VR [6, 7], 3DTV [8, 9] and holographic [10]–[12] systems. Capturing a DSLF of real scenes with wide field of view (FoV) requires high number of densely-located cameras which is not practical. For such scenes, only Sparsely-Sampled Light Fields (SSLFs) can be captured in practice. SSLFs can have *small disparities* between neighboring views in the range of 8 pixels, *moderate disparities* in the range of 15-16 pixels, and *large disparities* in the range of 30-32 pixels [13]. Real-world wide-FoV DSLFs are typically reconstructed from SSLFs using computational imaging approaches [14]–[18].

Related work. Video frame interpolation methods can be adapted to solve the DSLF reconstruction problem because a 3D SSLF can be treated as a virtual video sequence. Niklaus *et al*. have proposed Separable Convolution (SepConv), a learning-based video frame synthesis method using spatially adaptive kernels [19]. Bao *et al*. have proposed a Depth-Aware video frame INterpolation (DAIN) algorithm that leverages optical flow, local interpolation kernels, depth maps and contextual features [20]. Xu *et al*. have improved the linear models adopted in Linear Video Interpolation (LVI) methods [21, 22] and proposed Quadratic Video Interpolation (QVI) approach considering the acceleration information in videos [23]–[25]. Gao and Koch have extended SepConv for DSLF reconstruction by proposing Parallax-Interpolation Adaptive Separable Convolution (PIASC) [26].

The aforementioned methods tend to fail for the case of DSLF reconstruction from *large-disparity* SSLFs of complex scenes. To address the problem, Gao *et al*. have developed a learning-based DSLF reconstruction method, termed Deep Residual Shearlet Transform (DRST) [27]. It replaces the original shearlet-domain interactive regularization [28]–[30] with a learned one. The method however can only handle EPIs with a small number of input views, *i.e*. three views. For more views it has to be applied recursively, which reduces the speed.

To tackle this problem, a novel self-supervised [31] DSLF reconstruction method, CycleST, is proposed in this paper. Along with the shearlet transform, it leverages cycle consistency [32], recently used for video frame interpolation [33, 34]. Specifically, since several DSLFs with different angular resolutions can be reconstructed from the same input SSLF, the cycle consistency is the technique that guarantees these DSLFs have similar reconstruction results *w.r.t.* the same angular positions.

We summarize the main contributions of this letter as below:

- The proposed CycleST fully leverages the Deep Neural Network (DNN) with cycle-consistency loss in shearlet domain to perform EPI inpainting in image domain for an input SSLF with an arbitrary number of views.
- CycleST is fully self-supervised and trained solely on synthetic SSLFs with *small disparity ranges* ($\leqslant$ 8 pixels);
- Experimental results on challenging real-world SSLFs with large disparity ranges (16 - 32 pixels) demonstrate the superiority of CycleST over ST for DSLF reconstruction in terms of accuracy and efficiency ($\geqslant$ 8.9x speedup).

II. Methodology

A. Preliminaries

1) Symbols and notations: The symbols and notations used in this letter are summarized in Table I. The target DSLF $\mathcal{D}$ to be reconstructed from the input SSLF $\mathcal{S}$ has the same spatial resolution ($m \times l$) but different angular resolution. The angular resolutions $\dot{n}$ and n of $\mathcal{D}$ and $\mathcal{S}$ are related through the

This work was supported by the project "Modeling and Visualization of Perceivable Light Fields" funded by Academy of Finland under grant No. 325530 and carried out with the support of Centre for Immersive Visual Technologies (CIVIT) research infrastructure, Tampere University, Finland.

The authors are with the Unit of Computing Sciences (CS), Faculty of Information Technology and Communication Sciences (ITC), Tampere University, 33014 Tampere, Finland (e-mail: {yuan.gao, robert.bregovic, atanas.gotchev}@tuni.fi).

Table I
SYMBOLS AND NOTATIONS.

Symbol	Name	Description
$m \times l$	Spatial resolution of $\mathcal{S}$, $\mathcal{S}'$ and $\mathcal{D}$	width $\times$ height
n, n', $\dot{n}$	Angular resolutions of $\mathcal{S}$, $\mathcal{S}'$ and $\mathcal{D}$	
$d^{\mathcal{S}}_{\min}$	Minimum disparity of $\mathcal{S}$	$d^{\mathcal{S}}_{\min} = \min\big(\mathrm{disp}(\mathcal{S})\big)$
$d^{\mathcal{S}}_{\max}$	Maximum disparity of $\mathcal{S}$	$d^{\mathcal{S}}_{\max} = \max\big(\mathrm{disp}(\mathcal{S})\big)$
$d^{\mathcal{S}}_{\mathrm{range}}$	Disparity range of $\mathcal{S}$	$d^{\mathcal{S}}_{\mathrm{range}} = \big(d^{\mathcal{S}}_{\max} - d^{\mathcal{S}}_{\min}\big)$
ε_i	Sparsely-sampled EPI (SSEPI)	$\varepsilon_i \in \mathbb{R}^{m\times n\times 3}$
ϵ_i	SSEPI (for training)	$\epsilon_i \in \mathbb{R}^{m\times n'\times 3}$
ζ_i	Densely-sampled EPI	$\zeta_i \in \mathbb{R}^{m\times \dot{n}\times 3}$
$\mathcal{S}$	Large-disparity-range SSLF ($16 < d^{\mathcal{S}}_{\mathrm{range}} \leqslant 32$ pixels)	$\mathcal{S} = \{\varepsilon_i \vert 1 \leqslant i \leqslant l\}$
$\mathcal{S}'$	Small-disparity-range SSLF (for training) ($1 < d^{\mathcal{S}'}_{\mathrm{range}} \leqslant 8$ pixels)	$\mathcal{S}' = \{\epsilon_i \vert 1 \leqslant i \leqslant l\}$
$\mathcal{D}$	Target DSLF to be reconstructed from $\mathcal{S}$ ($d^{\mathcal{D}}_{\mathrm{range}} \leqslant 1$ pixel)	$\mathcal{D} = \{\zeta_i \vert 1 \leqslant i \leqslant l\}$
τ	Sampling interval ($\mathcal{D} \to \mathcal{S}$)	$\tau = \frac{\dot{n}-1}{n-1} = 32$
τ'	Sampling interval ($\mathcal{D} \to \mathcal{S}'$)	$\tau' = \frac{\dot{n}-1}{n'-1} = 8$

sampling interval τ such that $\tau = \frac{\dot{n}-1}{n-1}$, $\tau \geqslant d^{\mathcal{S}}_{\mathrm{range}}$. We consider input $\mathcal{S}$ with *large disparity ranges* ($16 < d^{\mathcal{S}}_{\mathrm{range}} \leqslant 32$ pixels), hence $\tau = 32$. Both $\mathcal{S}$ and $\mathcal{D}$ can be regarded as sets of EPIs, *i.e.* $\mathcal{S} = \{\varepsilon_i \vert 1 \leqslant i \leqslant l\}$ and $\mathcal{D} = \{\zeta_i \vert 1 \leqslant i \leqslant l\}$. The EPIs ε_i and ζ_i have different heights because of the different angular resolutions. To train the CycleST network, we utilize available SSLF datasets with *small disparity ranges* ($\leqslant 8$ pixels). Let $\mathcal{S}' = \{\epsilon_i \vert 1 \leqslant i \leqslant l\}$ denote one of these training SSLFs, where $d^{\mathcal{S}'}_{\mathrm{range}} \leqslant 8$ pixels and $\epsilon_i \in \mathbb{R}^{m\times n'\times 3}$. The sampling interval from $\mathcal{D}$ to $\mathcal{S}'$ is represented by τ', where $\tau' \geqslant d^{\mathcal{S}'}_{\mathrm{range}}$.

2) Shearlet Transform (ST): ST has been originally proposed in [35]–[37] and extended for DSLF reconstruction in [28]–[30], where an elaborately-tailored shearlet system with ξ scales has been developed for the angular resolution enhancement of any input $\mathcal{S}$ with two requirements: (i) $d^{\mathcal{S}}_{\min} \geqslant 0$ and (ii) $d^{\mathcal{S}}_{\max} \leqslant \tau$. The number of scales, *i.e.* ξ, is determined by the sampling interval τ as $\xi = \lceil \log_2 \tau \rceil$. The constructed ξ-scale shearlet system is used by shearlet analysis transform $\mathcal{SH} : \mathbb{R}^{\gamma\times\gamma} \to \mathbb{R}^{\gamma\times\gamma\times\eta}$ and shearlet synthesis transform $\mathcal{SH}^* : \mathbb{R}^{\gamma\times\gamma\times\eta} \to \mathbb{R}^{\gamma\times\gamma}$, where $\gamma \times \gamma$ represents the size of a shearlet filter and $\eta = (2^{\xi+1} + \xi - 1)$ denotes the number of shearlets. For $\tau = 32$, $\xi = 5$ and $\eta = 68$. Moreover, as suggested in [28], for $\tau = 32$, a good choice for γ is $\gamma = 255$.

B. CycleST

To resolve the challenging DSLF reconstruction problem for *large-disparity-range* SSLFs using *small-disparity-range* SSLF training data only, we propose a novel self-supervised method that leverages a residual learning-based DNN with EPI-reconstruction and cycle-consistency losses to restore EPI coefficients in shearlet domain. The proposed approach consists of five steps, namely (1) pre-shearing, (2) random cropping, (3) remapping, (4) sparse regularization and (5) post-shearing. Steps (1) - (4) form the self-supervised training part of CycleST. Steps (1) and (3) - (5) constitute the prediction part of CycleST. The details of these five steps are described as following.

1) Pre-shearing: To satisfy the two requirements of the elaborately-tailored shearlet system explained in Section II-A2, a pre-shearing operation is designed to change the disparities

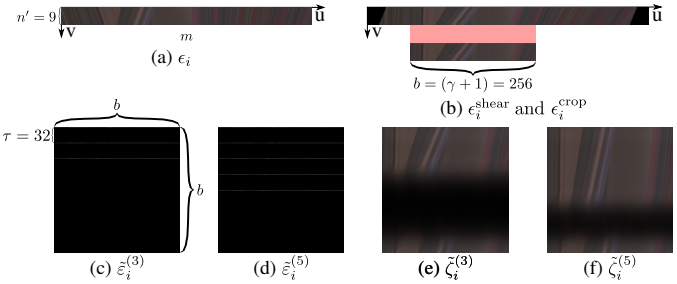

Figure 1. Illustration of the preparation of training data for CycleST. An EPI ϵ_i of the training SSLF $\mathcal{S}'$ is presented in (a). In (b), the top row shows $\epsilon_i^{\mathrm{shear}}$, *i.e.* the result of the pre-shearing operation on ϵ_i, and the bottom row shows $\epsilon_i^{\mathrm{crop}}$, which is the result of the random cropping step on $\epsilon_i^{\mathrm{shear}}$. The remapped EPIs $\tilde{\varepsilon}_i^{(3)}$ and $\tilde{\varepsilon}_i^{(5)}$ from $\epsilon_i^{\mathrm{crop}}$ are exhibited in (c) and (d), respectively. The corresponding reconstruction results of the sparse regularization step, *i.e.* $\tilde{\zeta}_i^{(3)}$ and $\tilde{\zeta}_i^{(5)}$, are displayed in (e) and (f), respectively.

of the input training SSLF $\mathcal{S}'$ using a shearing parameter $\rho^{\mathcal{S}'}$, where $\Big(d^{\mathcal{S}'}_{\min} - \big(\tau' - d^{\mathcal{S}'}_{\mathrm{range}}\big)\Big) \leqslant \rho^{\mathcal{S}'} \leqslant d^{\mathcal{S}'}_{\min}$. To be precise, each row v of ϵ_i is sheared by $(n' - v)\rho^{\mathcal{S}'}$ pixels, where $1 \leqslant v \leqslant n'$. One of the EPIs of the training SSLF $\mathcal{S}'$, *i.e.* ϵ_i, is displayed in Fig. 1 (a). The pre-sheared EPI corresponding to ϵ_i is represented by $\epsilon_i^{\mathrm{shear}}$ as illustrated in Fig. 1 (b).

2) Random cropping: To augment the number of training samples, a random cropping operation is leveraged to randomly cut an EPI $\epsilon_i^{\mathrm{crop}}$ from the above generated $\epsilon_i^{\mathrm{shear}}$ with a smaller width $b = (\gamma + 1) = 256$ pixels. Note that this operation does not crop any black border region of $\epsilon_i^{\mathrm{shear}}$. An example of the random cropping results is shown in Fig. 1 (b).

3) Remapping: To achieve self-supervision for CycleST, the rows of the cropped EPI $\epsilon_i^{\mathrm{crop}}$ are rearranged with zero-padding between neighboring rows, producing EPIs $\tilde{\varepsilon}_i^{(t)}$, *i.e.*

$$\tilde{\varepsilon}_i^{(t)}\Big(1 : \tau : \big((t-1)\tau + 1\big)\Big) = \epsilon_i^{\mathrm{crop}}\Big(1 : \tfrac{n'-1}{t-1} : n'\Big), \quad (1)$$

where $\tilde{\varepsilon}_i^{(t)} \in \mathbb{R}^{b\times b\times 3}$. As shown in Fig. 1 (c) and (d), two different EPIs, *i.e.* $\tilde{\varepsilon}_i^{(3)}$ and $\tilde{\varepsilon}_i^{(5)}$, are generated for each $\epsilon_i^{\mathrm{crop}}$, so that the cycle consistency information from them can be leveraged in the next sparse regularization step.

4) Sparse regularization: The remapped EPI $\tilde{\varepsilon}_i^{(t)}$ is then converted into shearlet coefficients via the shearlet analysis transform $\mathcal{SH}(\cdot)$. The sparse regularization step is essentially refining these coefficients in shearlet domain to fulfill image inpainting on $\tilde{\varepsilon}_i^{(t)}$ in image domain. To this end, a deep network using cycle-consistency loss with self-supervision setup is designed for the reconstruction of the shearlet coefficients.

Network architecture. As shown in Fig. 2, a residual convolutional neural network, based on the architectures of U-Net [39] and the generator network of CycleGAN [32], is adapted to perform the reconstruction of shearlet coefficients. The input data are $\tilde{\varepsilon}_i^{(t)}$, $t \in \{3, 5\}$, in image domain, which can be converted into coefficients $\mathcal{SH}\Big(\tilde{\varepsilon}_i^{(t)}\Big)$ with $\eta = 68$ channels in shearlet domain. These coefficients are then fed to the encoder-decoder network of CycleST, represented by $\mathcal{R}(\cdot)$, to predict the residuals for the input coefficients, *i.e.* $\mathcal{R}\Big(\mathcal{SH}\big(\tilde{\varepsilon}_i^{(t)}\big)\Big)$. The encoder component of $\mathcal{R}(\cdot)$ has four hierarchies, of which each is composed of one 3×3 convolution, one Leaky ReLU and one average pooling layers. The decoder part also has four hierarchies. Each one consists of the same convolution

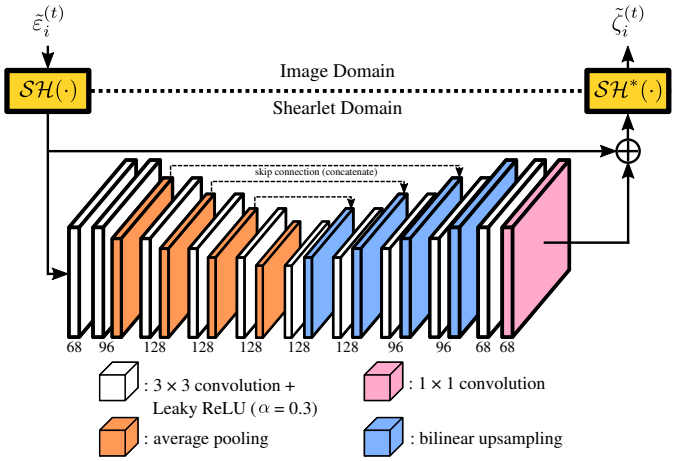

Figure 2. Architecture of the encoder-decoder network of CycleST, referred to as $\mathcal{R}(\cdot)$. The U-Net-based $\mathcal{R}(\cdot)$ and residual learning strategy [38] are utilized to refine the shearlet coefficients of the remapped EPI $\tilde{\varepsilon}_i^{(t)}$ in shearlet domain to produce an inpainted densely-sampled EPI $\tilde{\zeta}_i^{(t)}$ in image domain.

and Leaky ReLU layers as that of the encoder, but a bilinear upsampling layer instead of the pooling layer. The skip connections concatenate the outputs of last three hierarchies of both encoder and decoder. It can also be seen that the last layer of $\mathcal{R}(\cdot)$ is only a 1×1 convolution layer, without Leaky ReLU placed behind it. Following the residual learning strategy [38], the predicted coefficient residuals $\mathcal{R}\Big(\mathcal{SH}\big(\tilde{\varepsilon}_i^{(t)}\big)\Big)$ are merged with $\mathcal{SH}\Big(\tilde{\varepsilon}_i^{(t)}\Big)$ via an element-wise addition operation and then converted into reconstructed densely-sampled EPIs $\tilde{\zeta}_i^{(t)}$. For illustration, see Fig. 1 (e) and (f). Overall, the sparse regularization can be written as

$$\tilde{\zeta}_i^{(t)} = \mathcal{SH}^*\left(\mathcal{SH}\Big(\tilde{\varepsilon}_i^{(t)}\Big) + \mathcal{R}\Big(\mathcal{SH}\big(\tilde{\varepsilon}_i^{(t)}\big)\Big)\right). \tag{2}$$

Loss function. Two kinds of losses are considered in the loss function of CycleST network, *i.e.* the EPI-reconstruction loss $\mathcal{L}_{\mathrm{s}}(t)$ and cycle-consistency loss $\mathcal{L}_{\mathrm{cyc}}$. Both of them employ ℓ_1 norm, since recent research indicates that ℓ_1 norm is superior over ℓ_2 norm for learning-based view synthesis and image inpainting tasks [19, 40, 41]. The overall loss function $\mathcal{L}$ is a linear combination of $\mathcal{L}_{\mathrm{s}}(t)$, $t\in\{3,5\}$, and $\mathcal{L}_{\mathrm{cyc}}$, *i.e.*

$$\mathcal{L} = \mathcal{L}_{\mathrm{s}}(3) + \mathcal{L}_{\mathrm{s}}(5) + \lambda\mathcal{L}_{\mathrm{cyc}}\,. \tag{3}$$

The EPI-reconstruction loss $\mathcal{L}_{\mathrm{s}}(t)$ measures the reconstruction errors between the ground-truth sparsely-sampled $\epsilon_i^{\mathrm{crop}}$ and predicted densely-sampled $\tilde{\zeta}^{(t)}$ in a self-supervised manner:

$$\mathcal{L}_{\mathrm{s}}(t) = \left\|\tilde{\zeta}_i^{(t)}\Big(1:\tfrac{(t-1)\tau}{n'-1}:\big((t-1)\tau+1\big)\Big) - \epsilon_i^{\mathrm{crop}}\right\|_1. \tag{4}$$

The cycle-consistency loss $\mathcal{L}_{\mathrm{cyc}}$ calculates the reconstruction differences between the predicted $\tilde{\zeta}^{(3)}$ and $\tilde{\zeta}^{(5)}$:

$$\mathcal{L}_{\mathrm{cyc}} = \left\|\tilde{\zeta}_i^{(3)}\Big(1:(2\tau+1)\Big) - \tilde{\zeta}_i^{(5)}\Big(1:2:(4\tau+1)\Big)\right\|_1. \tag{5}$$

Finally, λ is empirically set to 2.

5) Post-shearing: The post-shearing operation is only used in the inference phase of CycleST. In terms of using CycleST for reconstructing $\mathcal{D}$ from $\mathcal{S}$, the input $\mathcal{S}$ has been sheared with parameter $\rho^{\mathcal{S}}$ in the pre-shearing stage. The post-shearing step compensates this through the same shearing operation on the top $\dot{n}$ rows of $\tilde{\zeta}_i^{(n)}$ with a new shearing parameter $-\frac{\rho^{\mathcal{S}}}{\tau}$. The post-sheared $\tilde{\zeta}_i^{(n)}$ is then cut by only keeping the top $\dot{n}$ rows of it to produce ζ_i of the target $\mathcal{D}$. It is suggested that $\rho^{\mathcal{S}} = d_{\min}^{\mathcal{S}}$.

Table II
DETAILS ABOUT THE TRAINING AND EVALUATION DATASETS.

3D light fields	m	l	n'	$\ddot{n}$	δ	n	$\dot{n}$
$\mathcal{S}'_j,\ 1\leqslant j\leqslant 78$	512	512	9	-	-	-	-
$\Psi_j^1,\ 1\leqslant j\leqslant 9$	1280	720	-	97	16	7	193
$\Psi_j^2,\ j\in\{1,2\}$	960	720	-	97	16	7	193
$\Psi_j^2,\ j\in\{3,4\}$	960	720	-	97	32	4	97

III. EXPERIMENTS

A. Experimental Settings

1) Training dataset: The Inria synthetic light field datasets contains 39 synthetic 4D SSLFs with disparities from -4 to 4 pixels [42]. These 4D SSLFs have the same angular resolution 9×9 and spatial resolution 512×512 pixels. For training, we only pick the 5-th row and 5-th column 3D SSLFs from each synthetic 4D SSLF. Therefore, the training data of CycleST consists of $\mathcal{S}'_j$, $1\leqslant j\leqslant 78$, $l=512$, $m=512$ and $n'=9$. The pre-shearing operation in Section II-B1 is repeated three times for each $\mathcal{S}'_j$ with different shearing parameters $\rho^{\mathcal{S}'_j}\in\left\{\left(d_{\min}^{\mathcal{S}'_j}-\left(\tau'-d_{\mathrm{range}}^{\mathcal{S}'_j}\right)\right),\left(d_{\min}^{\mathcal{S}'_j}-0.5\cdot\left(\tau'-d_{\mathrm{range}}^{\mathcal{S}'_j}\right)\right),d_{\min}^{\mathcal{S}'_j}\right\}$. As a result, the number of the generated $\epsilon_i^{\mathrm{crop}}$ for each training epoch is $78\times l=39,936$.

2) Evaluation Datasets: For evaluation, we consider two datasets of light fields with *tiny disparity ranges* ($\leqslant 2$ pixels). Though they are not DSLFs, they are otherwise suitable as they have a wide FoV and wide baseline. The Evaluation Dataset 1 (ED1) is the tailored High Density Camera Array (HDCA) dataset [43] using the same cutting and scaling strategy as in [27]. Consequently, nine *tiny-disparity-range* light fields Ψ_j^1, $1\leqslant j\leqslant 9$, form ED1 with the same spatial resolution ($m\times l=1280\times720$ pixels) and angular resolution ($\ddot{n}=97$). For each Ψ_j^1, an input SSLF $\mathcal{S}_j^1$ can be produced using an decimation rate $\delta=16$. As a result, the angular resolution of $\mathcal{S}_j^1$ is $n=\left(\frac{\ddot{n}-1}{\delta}+1\right)=7$. The target $\mathcal{D}_j^1$ to be reconstructed from $\mathcal{S}_j^1$ has angular resolution $\dot{n}=193$. The MPI light field archive contains two *tiny-disparity-range* light fields ("bikes" and "workshop") and two DSLFs ("mannequin" and "living room") [44], which constitute the Evaluation Dataset 2 (ED2), *i.e.* Ψ_j^2, $1\leqslant j\leqslant 4$. The spatial and angular resolutions of each Ψ_j^2 is $m=960$, $l=720$ and $\ddot{n}=97$. For the *tiny-disparity-range* Ψ_j^2, $j\in\{1,2\}$, the decimation rate δ is set to 16, such that $n=7$ and $\dot{n}=193$. For the densely-sampled Ψ_j^2, $j\in\{3,4\}$, the decimation rate δ is set to 32, such that $n=4$ and $\dot{n}=97$. The angular and spatial resolutions of the training and evaluation datasets are also summarized in Table II. The minimum disparity and disparity range of $\mathcal{S}_j^1$ and $\mathcal{S}_j^2$ are exhibited in Table III and Table IV, respectively.

3) Implementation details: The weights of all the filters of the CycleST network $\mathcal{R}(\cdot)$ are initialized by means of the He normal initializer [45]. The AdaMax optimizer [46] is employed to train the model for 12 epochs on an Nvidia GeForce RTX 2080 Ti GPU for around 33 hours. The learning rate is gradually reduced from 10^{-3} to 10^{-5} using an exponential decay rate during the first four epochs and then fixed to 10^{-5}

Table III
DISPARITY ESTIMATION, MINIMUM AND AVERAGE PER-VIEW PSNR RESULTS (IN DB) FOR THE PERFORMANCE EVALUATION OF DIFFERENT LIGHT FIELD RECONSTRUCTION METHODS ON ED1. THE BEST TWO RESULTS ARE HIGHLIGHTED IN RED AND BLUE COLORS.

j	**Disparity (pix)**		**Minimum PSNR / Average PSNR (dB)**							
	$d_{\min}^{\mathcal{S}_j^1}$	$d_{\text{range}}^{\mathcal{S}_j^1}$	SepConv ($\mathcal{L}_1$) [19]	PIASC ($\mathcal{L}_1$) [26]	DAIN [20]	LVI [23]	QVI [23]	ST [29]	DRST [27]	CycleST
1	25	19	19.988 / 21.769	19.978 / 21.760	29.042 / 32.664	32.382 / 33.397	32.164 / 33.475	32.133 / 35.185	32.452 / 35.027	34.288 / 35.918
2	27	22	20.777 / 23.978	20.782 / 24.015	22.563 / 24.516	24.828 / 26.073	24.854 / 26.090	25.877 / 27.953	23.811 / 25.512	25.409 / 26.712
3	28	27	24.081 / 26.969	24.089 / 27.013	25.077 / 27.794	27.940 / 29.660	28.292 / 29.724	26.672 / 29.403	26.725 / 28.622	28.614 / 30.841
4	25	30	24.648 / 28.486	24.660 / 28.584	27.125 / 28.765	28.482 / 30.225	28.552 / 30.115	29.153 / 32.639	29.162 / 31.179	29.320 / 31.470
5	25	29	26.942 / 29.060	26.954 / 29.135	28.330 / 29.739	30.129 / 31.095	30.361 / 31.173	30.780 / 33.111	30.737 / 31.637	31.177 / 31.913
6	25	29	26.965 / 29.620	26.977 / 29.692	31.003 / 34.817	32.588 / 34.198	31.796 / 34.126	33.853 / 36.354	34.118 / 36.712	36.006 / 37.513
7	26	17	21.223 / 24.750	21.224 / 24.784	22.645 / 24.718	24.488 / 26.202	24.760 / 26.252	25.458 / 27.876	24.458 / 26.423	25.428 / 27.024
8	28	21	21.152 / 24.309	21.158 / 24.360	22.320 / 24.633	24.627 / 26.122	24.724 / 25.974	26.137 / 28.451	24.500 / 26.549	26.301 / 28.046
9	28	27	26.455 / 29.750	26.468 / 29.839	26.791 / 30.658	30.451 / 31.636	30.829 / 32.069	29.721 / 32.252	29.169 / 31.513	31.745 / 33.963

Table IV
DISPARITY ESTIMATION, MINIMUM AND AVERAGE PER-VIEW PSNR RESULTS (IN DB) FOR THE PERFORMANCE EVALUATION OF DIFFERENT LIGHT FIELD RECONSTRUCTION METHODS ON ED2. THE BEST TWO RESULTS ARE HIGHLIGHTED IN RED AND BLUE COLORS.

j	**Disparity (pix)**		**Minimum PSNR / Average PSNR (dB)**							
	$d_{\min}^{\mathcal{S}_j^2}$	$d_{\text{range}}^{\mathcal{S}_j^2}$	SepConv ($\mathcal{L}_1$) [19]	PIASC ($\mathcal{L}_1$) [26]	DAIN [20]	LVI [23]	QVI [23]	ST [29]	DRST [27]	CycleST
1	-14	23.5	30.611 / 32.994	30.845 / 33.012	29.625 / 31.032	29.449 / 31.470	29.752 / 31.548	29.932 / 32.804	29.775 / 31.712	29.845 / 31.645
2	-6.5	23	34.155 / 37.138	34.324 / 37.363	33.186 / 34.341	34.013 / 35.254	34.300 / 35.410	33.911 / 37.286	34.107 / 35.887	34.773 / 36.138
3	-15	29	31.571 / 34.117	31.662 / 34.290	31.789 / 32.964	30.806 / 32.710	31.071 / 33.008	30.849 / 33.610	31.513 / 33.775	31.453 / 33.615
4	-12	28	37.106 / 41.760	37.371 / 42.797	37.198 / 40.341	36.849 / 39.368	36.793 / 39.616	36.069 / 40.104	36.444 / 40.415	36.924 / 40.610

for the rest eight epochs. The mini-batch for each training step is composed of two different samples ϵ_i^{crop} shown in Fig. 1 (b). The number of the trainable parameters of $\mathcal{R}(\cdot)$ is around 1.4 M. The implementation of ST is from [47].

B. Results and analysis

The proposed CycleST method is compared with the state-of-the-art video frame interpolation approaches, *i.e.* SepConv ($\mathcal{L}_1$) [19], DAIN [20], LVI [23], QVI [23], and DSLF reconstruction methods, *i.e.* PIASC ($\mathcal{L}_1$) [26], ST [29], DRST [27]. The performance of all the algorithms is compared using their minimum and average per-view PSNR values on each evaluation light field. The performance results on ED1 are exhibited in Table III. It can be seen from this table that all the results of CycleST are either the best or the second best, suggesting that the proposed method can effectively handle the DSLF reconstruction on SSLFs with repetitive patterns and large disparity ranges present in ED1. Specifically, the scenes of ED2 have no repetitive-pattern objects and less occlusions compared with ED1. The DSLF reconstruction results of all the methods on ED2 are presented in Table IV. As can be seen from the table, CycleST achieves the best minimum PSNR result on Ψ_2^2. In addition, it can also be seen that all the results of PIASC are either the best or the second best; however, in Table III, the results of PIASC and SepConv on ED1 are significantly worse than the other baseline approaches. This implies that the kernel-based PIASC and SepConv are not as robust as CycleST for DSLF reconstruction, since they may fail in DSLF reconstruction on large-disparity-range SSLFs with repetitive patterns and complex occlusions. Moreover, the proposed CycleST method outperforms the flow-based QVI and LVI on ED1 and ED2 *w.r.t.* minimum and average PSNRs. Finally, since CycleST and DRST are developed based on ST, the computation time of all these three methods is compared in Table V for the same computing platform. It can be seen that CycleST is at least 8.9 times faster than ST and 2.1 times faster than DRST.

Table V
THE AVERAGE COMPUTATION TIME AND SPEEDUP OVER ST OF RECONSTRUCTING ζ_i FROM A SPARSELY-SAMPLED $\varepsilon_i \in \mathbb{R}^{m\times n\times 3}$.

m	n	ST [29]	DRST [27]	CycleST
1280	7	3324 ms (1.0 x)	742 ms (4.5 x)	275 ms (12.1 x)
960	7	2257 ms (1.0 x)	598 ms (3.8 x)	253 ms (8.9 x)
960	4	1792 ms (1.0 x)	365 ms (4.9 x)	177 ms (10.1 x)

IV. CONCLUSION

This letter has presented a novel self-supervised DSLF reconstruction method, CycleST, which refines the shearlet coefficients of the densely-sampled EPIs in shearlet domain to perform the inpainting of them in image domain. The proposed CycleST takes full advantage of the shearlet transform, encoder-decoder network with residual learning strategy and two types of loss functions, *i.e.* the EPI-reconstruction and cycle-consistency losses. Besides, CycleST is trained in a self-supervised fashion solely on synthetic SSLFs with small disparity ranges. Experimental results on two real-world evaluation datasets demonstrate that CycleST is extremely effective for DSLF reconstruction on SSLFs with large disparity ranges (16 - 32 pixels), complex occlusions and repetitive patterns. Moreover, CycleST achieves $\sim$9x speedup over ST, at least.

REFERENCES

[1] M. Levoy and P. Hanrahan, "Light field rendering," in *ACM SIGGRAPH*, 1996, pp. 31–42.

[2] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *ACM SIGGRAPH*, 1996, pp. 43–54.

[3] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Densely-sampled light field reconstruction," in *Real VR – Immersive Digital Reality: How to Import the Real World into Head-Mounted Immersive Displays*. Springer International Publishing, 2020, pp. 67–95.

[4] R. Bregovic, E. Sahin, S. Vagharshakyan, and A. Gotchev, "Signal processing methods for light field displays," in *Handbook of Signal Processing Systems*. Springer International Publishing, 2019, pp. 3–50.

[5] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," *IEEE JSTSP*, vol. 11, no. 7, pp. 926–954, 2017.

[6] R. S. Overbeck, D. Erickson, D. Evangelakos, M. Pharr, and P. Debevec, "A system for acquiring, processing, and rendering panoramic light field stills for virtual reality," *ACM TOG*, vol. 37, no. 6, pp. 1–15, 2018.

[7] J. Yu, "A light-field journey to virtual reality," *IEEE MultiMedia*, vol. 24, no. 2, pp. 104–112, 2017.

[8] A. Smolic, "3D video and free viewpoint video - from capture to display," *Pattern Recognition*, vol. 44, no. 9, pp. 1958–1968, 2011.

[9] M. Tanimoto, M. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint TV," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 67–76, 2011.

[10] M. Yamaguchi, "Light-field and holographic three-dimensional displays," *JOSA A*, vol. 33, no. 12, pp. 2348–2364, 2016.

[11] T. Balogh, "The HoloVizio system," in *SPIE Stereoscopic Displays and Virtual Reality Systems XIII*, vol. 6055, 2006, pp. 279 – 290.

[12] T. Agocs, T. Balogh, T. Forgacs, F. Bettio, E. Gobbetti, G. Zanetti, and E. Bouvier, "A large scale interactive holographic display," in *IEEE VR*, 2006, pp. 311–311.

[13] S. Vagharshakyan, O. Suominen, R. Bregovic, and A. Gotchev, "ICME 2018 grand challenge on densely sampled light field reconstruction," https://civit.fi/icme-2018-grand-challenge/, 2018.

[14] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[15] M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. DuVall, J. Dourgarian, J. Busch, M. Whalen, and P. Debevec, "Immersive light field video with a layered mesh representation," *ACM TOG*, vol. 39, no. 4, pp. 86:1–86:15, 2020.

[16] J. Shi, X. Jiang, and C. Guillemot, "Learning fused pixel and feature-based view reconstructions for light fields," in *CVPR*, 2020, pp. 2555–2564.

[17] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM TOG*, vol. 38, no. 4, pp. 1–14, 2019.

[18] J. Flynn, M. Broxton, P. Debevec, M. DuVall, G. Fyffe, R. Overbeck, N. Snavely, and R. Tucker, "Deepview: View synthesis with learned gradient descent," in *CVPR*, 2019, pp. 2367–2376.

[19] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *ICCV*, 2017, pp. 261–270.

[20] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in *CVPR*, 2019, pp. 3703–3712.

[21] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High quality estimation of multiple intermediate frames for video interpolation," in *CVPR*, 2018, pp. 9000–9008.

[22] Z. Liu, R. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *ICCV*, 2017, pp. 4473–4481.

[23] X. Xu, S. Li, W. Sun, Q. Yin, and M.-H. Yang, "Quadratic video interpolation," in *NeurIPS*, 2019, pp. 1645–1654.

[24] S. Li, X. Xu, Z. Pan, and W. Sun, "Quadratic video interpolation for VTSR challenge," in *ICCV Workshops*, 2019, pp. 3427–3431.

[25] S. Nah, S. Son, R. Timofte, K. M. Lee *et al.*, "AIM 2019 challenge on video temporal super-resolution: methods and results," in *ICCV Workshops*, 2019, pp. 3388–3398.

[26] Y. Gao and R. Koch, "Parallax view generation for static scenes using parallax-interpolation adaptive separable convolution," in *ICME Workshops*, 2018, pp. 1–4.

[27] Y. Gao, R. Bregovic, R. Koch, and A. Gotchev, "DRST: Deep residual shearlet transform for densely-sampled light field reconstruction," *arXiv preprint arXiv:2003.08865*, 2020.

[28] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform," *IEEE TPAMI*, vol. 40, no. 1, pp. 133–147, 2018.

[29] ——, "Accelerated shearlet-domain light field reconstruction," *IEEE JSTSP*, vol. 11, no. 7, pp. 1082–1091, 2017.

[30] ——, "Image based rendering technique via sparse representation in shearlet domain," in *ICIP*, 2015, pp. 1379–1383.

[31] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE TPAMI*, 2020.

[32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.

[33] F. A. Reda, D. Sun, A. Dundar, M. Shoeybi, G. Liu, K. J. Shih, A. Tao, J. Kautz, and B. Catanzaro, "Unsupervised video interpolation using cycle consistency," in *ICCV*, 2019, pp. 892–900.

[34] Y.-L. Liu, Y.-T. Liao, Y.-Y. Lin, and Y.-Y. Chuang, "Deep video frame interpolation using cyclic frame generation," in *AAAI*, vol. 33, 2019, pp. 8794–8802.

[35] G. Kutyniok, W.-Q. Lim, and R. Reisenhofer, "Shearlab 3D: Faithful digital shearlet transforms based on compactly supported shearlets," *ACM Transactions on Mathematical Software (TOMS)*, vol. 42, no. 1, 2016.

[36] G. Kutyniok, M. Shahram, and X. Zhuang, "Shearlab: A rational design of a digital parabolic scaling algorithm," *SIAM Journal on Imaging Sciences*, vol. 5, no. 4, pp. 1291–1332, 2012.

[37] G. Kutyniok and D. Labate, *Shearlets: Multiscale analysis for multivariate data*. Springer Science+Business Media, 2012.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[40] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *CVPR*, 2017, pp. 2270–2279.

[41] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *ICCV*, 2019, pp. 4471–4480.

[42] J. Shi, X. Jiang, and C. Guillemot, "A framework for learning depth from a flexible subset of dense and sparse light field views," *IEEE TIP*, vol. 28, no. 12, pp. 5867–5880, 2019.

[43] M. Ziegler, R. op het Veld, J. Keinert, and F. Zilly, "Acquisition system for dense lightfield of large scenes," in *3DTV-CON*, 2017, pp. 1–4.

[44] V. K. Adhikarla, M. Vinkler, D. Sumin, R. K. Mantiuk, K. Myszkowski, H.-P. Seidel, and P. Didyk, "Towards a quality metric for dense light fields," in *CVPR*, 2017, pp. 58–67.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015, pp. 1026–1034.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[47] Y. Gao, R. Koch, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform in tensorflow," in *ICME Workshops*, 2019, pp. 612–612.