Self-supervised Light Field View Synthesis Using Cycle Consistency

Yang Chen, Martin Alain, Aljosa Smolic

V-SENSE project, School of Computer Science and Statistics, Trinity College Dublin

{cheny5, alainm, smolica}@scss.tcd.ie

Abstract—High angular resolution is advantageous for practical applications of light fields. In order to enhance the angular resolution of light fields, view synthesis methods can be utilized to generate dense intermediate views from sparse light field input. Most successful view synthesis methods are learning-based approaches which require a large amount of training data paired with ground truth. However, collecting such large datasets for light fields is challenging compared to natural images or videos. To tackle this problem, we propose a self-supervised light field view synthesis framework with cycle consistency. The proposed method aims to transfer prior knowledge learned from high quality natural video datasets to the light field view synthesis task, which reduces the need for labeled light field data. A cycle consistency constraint is used to build bidirectional mapping enforcing the generated views to be consistent with the input views. Derived from this key concept, two loss functions, cycle loss and reconstruction loss, are used to fine-tune the pre-trained model of a state-of-the-art video interpolation method. The proposed method is evaluated on various datasets to validate its robustness, and results show it not only achieves competitive performance compared to supervised fine-tuning, but also outperforms state-of-the-art light field view synthesis methods, especially when generating multiple intermediate views. Besides, our generic light field view synthesis framework can be adopted to any pre-trained model for advanced video interpolation.

Index Terms—Light Fields View Synthesis, Video Interpolation, Cycle Consistency, Self-supervised Fine-tuning

I. INTRODUCTION

Light field imaging has been introduced to the computer graphics and computer vision community over 20 years ago in the pioneer work of Levoy et al. [1], and has gained a lot of attention since. Compared to traditional 2D imaging systems, 4D light field imaging systems aim to collect all light rays passing through a given 3D volume by capturing not only two spatial but also two additional angular dimensions. Light field imaging has many applications ranging from post-capture photograph refocusing to virtual and augmented reality [2].

Advantages of dense light fields have been demonstrated for various computer vision and graphics tasks compared to sparse light fields, including depth estimation, object segmentation and image-based rendering [3]. However, a trade-off has usually to be made between spatial and angular resolution when implementing light field capture systems. While lenslet plenoptic cameras tend to favor angular resolution over spatial

resolution [4], modern camera arrays allowing light field video capture have physically limited angular sampling but can capture high resolution 2D views [5].

Thus, light field resolution enhancement has been a very active research topic. While enhancing the spatial resolution can be seen as an extension of 2D image super-resolution methods [6], [7], enhancing the angular resolution has been explored from various angles, such as angular super-resolution, view synthesis or interpolation, light field reconstruction, or epipolar-plane image (EPI) inpainting [7]–[11]. All these methods eventually generate novel intermediate views from a sparse set of input views and thus obtain a denser light field.

The study of these angular resolution enhancement approaches shows that learning-based methods achieve the best performance, and motivates the work presented in this paper. One of the first learning-based approaches for light field view synthesis was introduced in [9], and proposed to train two convolutional neural networks (CNN) first estimating the disparity, and then using this disparity to synthesize the intermediate views. It has been observed that the method proposed by Kalantari et.al. [9] can fail when dealing with complicated scenes containing occluded regions, non-Lambertian areas and large displacements. A "blur-detail restoration-deblur" framework was then proposed to enhance light field angular resolution using EPIs in [10]. An end-to-end network with 3D detail on pseudo 4D EPI was then introduced in [11]. These generative methods are prone to over-smoothing for large displacements that leads to a loss in perceptual quality. The limitations of these learning-based methods can partly be explained by the limited size of the training datasets.

As mentioned above, implementing real-world light field capture systems is a complex task, which inherently limits the size of available datasets, and can only be partly compensated by the use of synthetic datasets. In comparison, a huge amount of real-world natural images and videos are readily available, which allows to train powerful learning-based methods, and it has been shown recently that a direct extension of single image learning-based methods can outperform light field learning-based approaches for spatial super-resolution [12].

To address the aforementioned issues, we propose a self-supervised learning-based light field view synthesis framework based on existing video view synthesis methods which benefit from large real-world training datasets.

Rather than training a leaning-based method from scratch, the proposed framework adapts the existing video view synthe-

sis method to light fields using fine-tuning only, which requires smaller training datasets. Cycle consistency has proven capable of modeling invertible mapping when direct supervision is unavailable [13]–[15]. Thus, a cycle consistency constraint is introduced in our framework to allow self-supervised training without the need of paired ground truth, which further reduces the size of training data required.

In summary, our main contributions in this paper are:

- We introduce a novel light field view synthesis framework utilizing natural priors from existing large image datasets.
- We propose for the first time a self-supervised fine-tuning training approach for light fields based on the cycle consistency constraint.
- We demonstrate that the proposed framework outperforms state-of-the-art light field view synthesis methods.

This paper is organized as follows. In Section II, we review existing related work about light field angular super-resolution, general video frame interpolation and applications of the cycle consistency. In Section III, the proposed light field view synthesis framework and related self-supervised training details are explained. Then, our proposed method is evaluated on various datasets and compared to state-of-the-art light field view synthesis methods in Section IV. Finally, we present our conclusions in Section V.

II. RELATED WORK

A. Light Field View Synthesis

1) Optimization-based methods.: Dense light fields are very sparse in the transform domain, which can be used as a powerful prior. Therefore, Shi et al. proposed an optimization framework in the continuous Fourier domain to reconstruct the dense light field [16]. A more advanced framework using the shearlet transform is proposed in [8] which performs inpainting on the EPIs to recover the missing views. While these methods can achieve competitive performance, they require specific inputs which are not flexible and can make them difficult to use in practice.

2) Learning-based methods.: Kalantari et al. proposed a learning based framework for light field view synthesis [9] which estimates with a first CNN the disparity from 4 corner input views of a dense light field. A second CNN is then used to synthesize the target intermediate views using both the disparity and the 4 input views. Wu et al. re-modeled light field angular super-resolution as a detail restoration problem in the 2D EPI space [10]. A detail restoration framework is built to process EPIs of a sparse light field and to recover the angular details with a CNN. To exploit the inherent consistency of the light field, Wang et al. introduced an end-to-end network with pseudo 4D convolution by combining a 2D convolution on EPIs and a sequential 3D convolution [11]. Yeung et al. [17] also propose a two-step method, which first generates the whole set of novel views using a view synthesis network, and then retrieves texture details using a view refinement network.

Zhou et al. [18] train a deep network to predict the Multi-Plane Images (MPI) representation from a narrow-baseline
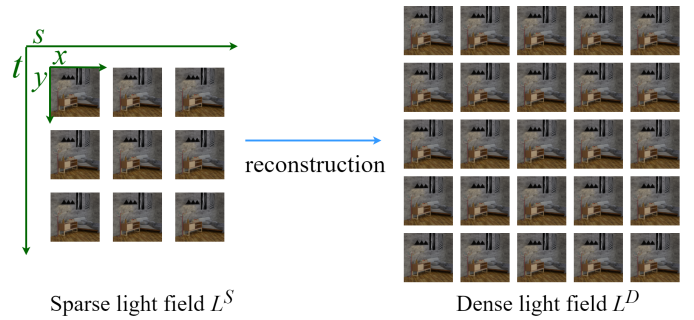

Fig. 1: Dense light field reconstruction. We aim to reconstruct a dense light field L^D with angular resolution (N, N) from a sparse light field L^S with angular resolution (n, n). The spatial resolution (h, w) of each view remains unchanged.

stereo image pair. The MPI representation can be use to generate novel views using homographies to reproject each plane of the MPI to the desired viewport. Mildenhall et al. [19] extend this method to light fields by promoting an MPI to each view of the light field. Novel views are then synthesized by combining intermediate generated views from the closest MPIs.

Video interpolation has also been applied to light field view synthesis in [20], using fully supervised fine-tuning with conventional loss functions on a small light field dataset. Although various existing methods have been proposed to enhance the angular resolution with impressive results, the robustness and generalization of these methods are limited by the quantity and quality of available real-world light field datasets, as explained in the previous section.

B. Video Frame Interpolation

Niklaus et al. proposed to estimate motion and color interpolation within one stage using an adaptive 2D kernel which is estimated from a trained convolutional neural network [21]. However, 2D kernel estimation requires huge memory to store information for all pixels and this shortcoming is addressed by replacing the 2D kernel with two separable 1D convolutional kernels [22].

C. Cycle Consistency

The key element of our proposed method is the introduction of the cycle consistency to the light field angular domain. The cycle consistency constraint aims to regularize structured predictions and establishes bidirectional mapping instead of unidirectional mapping built by conventional cost measurement functions.

Our work is inspired by the success of cyclic image generation for video interpolation [14], [15], which demonstrated the strength of cycle consistency to adapt a pre-trained model to a new target domain.

III. LIGHT FIELD VIEW SYNTHESIS USING CYCLE CONSISTENCY

A. Problem Formulation

In this work, a 4D light field L is parameterized using the two parallel planes representation as depicted in Figure 1,
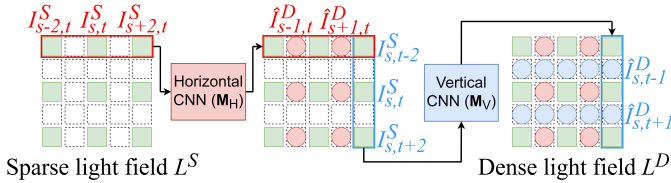
Fig. 2: An overview of our proposed view interpolation approach. Horizontal and vertical interpolation are cascaded to reconstruct L^D from L^S using view triplets from the corresponding angular dimensions.
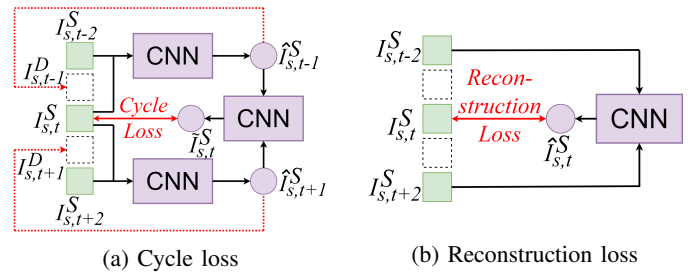

(a) Cycle loss   (b) Reconstruction loss

Fig. 3: Illustration of the cycle loss and reconstruction loss on a vertical input triplet. Both losses do not require any knowledge of the ground truth intermediate views (represented by dashed square) and can therefore be used for self-supervised training.

indexed by x, y over the spatial dimensions and s, t are the angular dimensions. We denote by $I_{s,t}$ the view extracted from a light field L at the angular position s,t. Given a sparsely-sampled light field L^S with resolution $(h \times w \times n \times n)$, the goal is to reconstruct a more densely-sampled light field L^D with the same spatial resolution and a higher angular resolution $(h \times w \times N \times N)$, where $N = \alpha(n-1) + 1$ and α is the up-sampling factor in the angular domain. Unless mentioned specifically, $\alpha = 2$ is used as default to explain our method. By fixing one angular dimension, a set of views can be extracted along the remaining angular dimension of the light field. Such a view set can be considered as a consecutive frame sequence, which can be captured by a virtual camera moving along the corresponded angular direction. Thus, the dense light field reconstruction problem can be treated as a video interpolation process along the fixed angular dimension. As listed in section II, many CNN-based methods have been shown to be successful for video interpolation tasks. However, directly adopting a pre-trained network from an existing video interpolation method to the light field domain may fail since the distribution of these two kind of data may differ. On the other hand, retraining a CNN from scratch can be laborious and the limited size of light field datasets may not allow to reach competitive performance. Thus, to maximally leverage the advantage of the cutting-edge video interpolation methods and to avoid its troublesome retraining, we introduce a self-supervised fine-tuning approach using cycle consistency to apply the pre-trained model of a video interpolation method to the light field domain.

B. Proposed Framework with Self-Supervised Learning

Given a sparse light field L^S with angular resolution (n, n), our proposed approach aims to build a learning based model that takes this light field as input and accurately reconstructs a high quality dense light field L^D with angular resolution (N, N) without the support of paired ground truth. As shown in Figure 2, we consider triplets of views extracted from L^S along a fixed angular dimension, either horizontally $\{I^S_{s-2,t}, I^S_{s,t}, I^S_{s+2,t}\}$ or vertically $\{I^S_{s,t-2}, I^S_{s,t}, I^S_{s,t+2}\}$. Note that triplets have to be used due to our proposed cycle loss described below. A dense light field L^D is obtained by first performing horizontal interpolation on all rows, and then performing vertical interpolation on all columns. The view

interpolation is achieved by two CNNs which share the same architecture but are trained separately along the horizontal and vertical dimensions.

Let us consider the horizontal interpolation case in order to explain the framework more in detail. Given an input triplet $\{I^S_{s-2,t}, I^S_{s,t}, I^S_{s+2,t}\}$, two intermediate views can be generated from pairwise adjacent views:

$$
\begin{aligned}
\hat{I}^D_{s-1,t} &= \mathbf{M}(I^S_{s-2,t}, I^S_{s,t}) \\
\hat{I}^D_{s+1,t} &= \mathbf{M}(I^S_{s,t}, I^S_{s+2,t})
\end{aligned}
\tag{1}
$$

where $\mathbf{M}$ is a pre-trained video interpolation method.

Inspired by the recent success of the application of the cycle consistency for video interpolation [14], [15], we propose to fine-tune our baseline interpolator $\mathbf{M}$ in a self-supervised manner by applying the cycle consistency constraint to the light field angular domain, as shown in Figure 3a. By applying the interpolator $\mathbf{M}$ on the two intermediate views generated from the input triplet as defined in equation 1, we can obtain an estimate of the center view of the input triplet $I^S_{s,t}$ which we denote as the cycle-reconstructed view $\tilde{I}^S_{s,t}$:

$$
\begin{aligned}
\tilde{I}^S_{s,t} &= \mathbf{M}\left(\hat{I}^D_{s-1,t},\ \hat{I}^D_{s+1,t}\right) \\
&= \mathbf{M}\left(\mathbf{M}(I^S_{s-2,t}, I^S_{s,t}),\ \mathbf{M}(I^S_{s,t}, I^S_{s+2,t})\right)
\end{aligned}
\tag{2}
$$

We can thus define the cycle-loss as the ℓ_1-norm distance between the cycle-reconstructed view $\hat{I}^S_{s,t}$ and the input view $I^S_{s,t}$:

$$
\mathcal{L}_c = ||\tilde{I}^S_{s,t} - I^S_{s,t}||_1
\tag{3}
$$

While ℓ_1-norm based losses are able to minimize the overall error between the estimated images and the corresponding original images, they are known to generate over-smooth results. To tackle this problem, we also introduce in our framework a perceptual loss $\mathcal{L}_p$, defined as the ℓ_2-norm between high-level convolutional features extracted from the cycle-reconstructed view and the input view:

$$
\mathcal{L}_p = ||\Psi(\tilde{I}^S_{s,t}) - \Psi(I^S_{s,t})||_2
\tag{4}
$$

where Ψ extracts the convolutional features from images using a VGG-16 network [23], which then is applied to train our base CNN network (SepConv [22], see below).
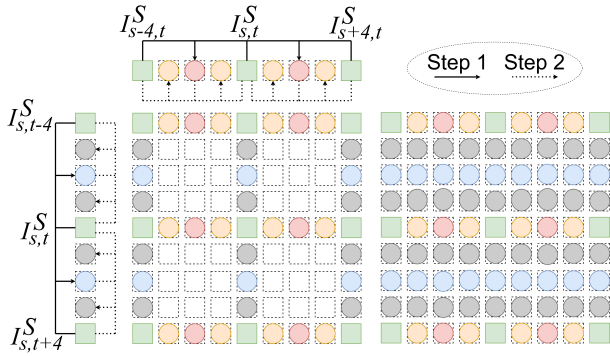
Fig. 4: Demonstration of the two-step strategy to generate multiple intermediate views from a set of sparse views, denoted as green squares, when $\alpha = 4$. The first step is to synthesize middle views, denoted as red and blue circles, between each pairwise input views. The second step uses original and synthetic views to reconstruct the remaining missing views, denoted as yellow and grey circles, along one angular dimension.

Furthermore, to stabilize the training process, we introduce a reconstruction loss $\mathcal{L}_r$, as shown in Figure 3b. In this case, the two non-adjacent views from the input triplet $I_{s-2,t}^S$ and $I_{s+2,t}^S$ are used to generate the center view of the input triplet:

$$\hat{I}_{s,t}^S = \mathbf{M}\left(I_{s-2,t}^S, I_{s+2,t}^S\right) \tag{5}$$

This reconstructed view can be used to define the reconstruction loss $\mathcal{L}_r$ as its ℓ_1-norm distance to the input view:

$$\mathcal{L}_r = ||\hat{I}_{s,t}^S - I_{s,t}^S||_1 \tag{6}$$

Note that all losses introduced in our framework as defined in equations 3, 4, and 6, do not rely on any knowledge of the ground truth dense light field L_D but only the given sparse input light field L_S, thus allowing to perform self-supervised training or fine-tuning of the learning-based interpolator $\mathbf{M}$.

Multi-step Light Field Generation While our framework naturally performs angular up-sampling with a factor $\alpha = 2$, denser light fields can be obtained by iteratively applying the proposed approach, as illustrated in Figure 4 for $\alpha = 4$. Any upsampling factor which is a power of two is in fact supported, i.e. $\alpha = 2^x, (x \in \mathbb{Z} \ \& \ x > 1)$.

C. Implementation Details

In this work, we select the adaptive separable convolution (SepConv) [22] as our baseline interpolator $\mathbf{M}$ due to its balance between ease of use and performance accuracy, but note that any learning-based video interpolation method [24]–[26] can be used within our framework. The network of SepConv employs an encoder-decoder architecture, each part contains convolution blocks and skip connections, to extract features and then performs four 1D kernel estimations individually to obtain the final results. We use the implementation available online based on *PyTorch* [12] and use the default configurations

[1] github.com/sniklaus/sepconv-slomo
[2] github.com/HyeongminLEE/pytorch-sepconv

from the original SepConv paper. We fine-tune the pre-trained model by minimizing the objective function:

$$\arg\min_{\mathbf{M}} \left(\lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r + \lambda_p \mathcal{L}_p\right) \tag{7}$$

where $\mathcal{L}_c$, $\mathcal{L}_r$ and $\mathcal{L}_p$ are defined in equations (3), (6) and (4). For all experiments, we set the parameters as $\lambda_c = 1$, $\lambda_r = 1$, and $\lambda_p = 0.06$. The Adam optimizer is applied for optimization with a batch size of 8. We start with the learning rate of 0.001 and a scheduler is applied to decay the rate according to the learning progress. As in the original SepConv work, we firstly crop training data to 150×150 patches, then randomly crop to 128×128. In addition, we perform pre-processing to eliminate patches containing too small disparity. An Intel Core i7-6700k 4.0GHz CPU was used for all our experiments, and the neural network training was run on a single Nvidia Titan Xp GPU with 12 GB memory.

IV. EXPERIMENTS

In this section, we first conduct an ablation study, especially evaluating the efficiency of the proposed framework compared to supervised fine-tuning. For this purpose, we use a variety of real-world and synthetic dense light field datasets which we sub-sample to create our test sparse datasets with sampling ratios $\alpha = 2$ and $\alpha = 4$.

We then compare the proposed framework to two top-performing state-of-the-art light field view synthesis methods, a shearlet-based method [8] and a learning-based method (LFEPICNN) [10].

For all our evaluations, the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) are computed over RGB images to evaluate the numerical performance of the different methods. For each light field, unless emphasized specifically, the average numerical results are computed over all synthesized views. All evaluations are performed on the same machine to ensure fairness of the comparison. Please find more detailed experimental summary on our website [3].

A. Ablation Study

For this study we used dense light fields from real-world and synthetic datasets. For the real-world dataset, we selected 27 real-world Lytro light fields captured by EPFL [27] and INRIA [28] using Lytro Illum cameras, and 11 light fields from the Stanford dataset taken by a camera gantry [29]. The Lytro Illum light fields are processed with the pipeline of Matysiak et al. [30]. For the synthetic light field dataset, all 28 light fields from the HCI benchmark [31] were used, as well as 160 light fields from the dataset of [32].

For testing, 10 light fields are used: 2 from EPFL, 2 from INRIA, 2 from Stanford, and 4 from HCI. All remaining light fields are used for training.

Test sparse light fields are sub-sampled from the original light fields with ratios $\alpha = 2$ and $\alpha = 4$. More precisely, 9×9 views are extracted from input light fields and considered as

[3] https://v-sense.scss.tcd.ie/?p=5163

dense ground truth, and 5×5 and 3×3 views are then sub-sampled to create sparse light fields.

We conduct the ablation experiments by comparing to several variants of the proposed framework. First, we use the pre-trained model of SepConv as the baseline. Since the dense light field ground truth is available, we fine-tuned the SepConv model using supervised training. We also evaluate the influence of the cycle loss by training our framework using only the reconstruction loss. We also assess the performance of our framework when vertical interpolation is performed before horizontal interpolation, as opposed applying horizontal interpolation first as shown in Figure 2.

The numerical results are computed and averaged over all test light fields, and the comparison is presented in Table I. As we can observe, our proposed method can outperform the pre-trained model even without the support of the ground-truth, and achieve competing performance compared to fully supervised fine-tuning. It is also clear that the use of the cycle loss improves the performance of our framework. In addition, we can see that cascading order of horizontal/vertical or vertical/horizontal interpolation has a non-negligible impact on the final performance.

TABLE I: Quantitative results of the ablation study.

	$\alpha = 2$		$\alpha = 4$	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
SepConv Pretrained	37.23	0.9880	34.66	0.9793
SepConv Supervised Fine-tuning	38.40	0.9921	35.81	0.9831
Ours without Cycle Loss	38.01	0.9883	35.25	0.9801
Ours with V-H CNN	38.14	0.9889	35.67	0.9817
Ours Full Model	38.30	0.9902	35.72	0.9830

B. Comparison to Light Field View Synthesis Methods

We use here the same test datasets as for the ablation study to compare the proposed framework against the pre-trained SepConv model, shearlet-based reconstruction [8], and LFEPICNN [10]. We used the implementations provided by authors, and carefully selected their parameters to maximize their performance.

We show the quantitative results for each dataset separately in Table II, III, and IV, as each dataset corresponds to a different disparity range. Note that the results are averaged per dataset.

The shearlet-based reconstruction is almost always outperformed by all other methods, and while shearlet-based reconstruction and LFEPICNN are designed specifically for light fields, they are only competitive on the Lytro dataset which has a very narrow disparity range. Our proposed framework consistently outperforms all other methods including SepConv. In addition, our method is more robust when using sparser input datasets such as when using a sub-sampling ratio $\alpha = 4$.

We present visual comparisons for the *ChezEdgar* and *LegoKnights* light fields in Figure 5. *LegoKnights* is a challenging case as it has wider disparity than other test light fields and large texture-less regions. Shearlet [8] and LFEPICNN [10] both fail to produce plausible results and significant artifacts

TABLE II: Numerical results on the real-world Lytro datasets [27], [28]

	$\alpha = 2$		$\alpha = 4$	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
Shearlet	33.10	0.9667	29.99	0.9361
LFEPICNN	35.35	0.9864	32.06	0.9640
SepConv	35.30	0.9836	32.46	0.9712
Ours	**36.76**	**0.9876**	**33.62**	**0.9767**

TABLE III: Numerical results on the synthetic HCI dataset [31]

	$\alpha = 2$		$\alpha = 4$	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
Shearlet	34.81	0.9734	29.88	0.8911
LFEPICNN	34.25	0.9692	30.42	0.9172
SepConv	38.88	0.9943	36.23	0.9888
Ours	**39.87**	**0.9953**	**37.44**	**0.9913**

TABLE IV: Numerical results on the real-world Stanford Gantry datasets [29]

	$\alpha = 2$		$\alpha = 4$	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
Shearlet	31.44	0.8977	29.03	0.8484
LFEPICNN	34.68	0.9407	30.46	0.8762
SepConv	37.80	0.9843	35.19	0.9788
Ours	**38.23**	**0.9853**	**36.47**	**0.9791**

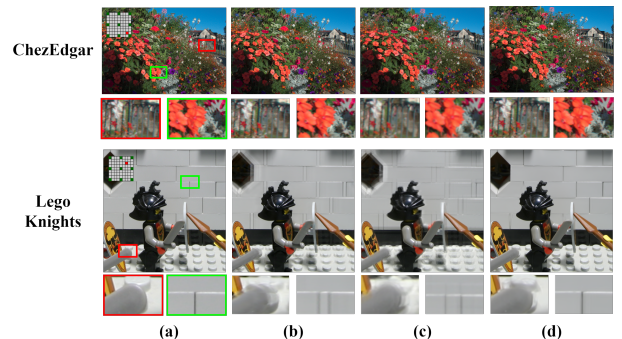
Fig. 5: Visual comparison on the INRIA ChezEdgar and Stanford Lego Knights light fields. (a) Ground-truth. (b) Shearlet [8]. (c) LFEPICNN [10]. (d) Ours

can be observed on challenging areas, such as the tip of the sword and bricks on the background wall. In comparison, our proposed approach generates results closer to the ground-truth. It demonstrates that our method is more robust to different real-world scenes and is able to produce more photo-realistic results for large disparity view synthesis.

A visual comparison of synthetic scenes is presented in Figure 6 using *Herbs* and *Bicycle* from the HCI dataset [31]. As we can observe, Shearlet [8] fails to reconstruct sharp details in texture-less regions, such as the door in *Bicycle*. The results of LFEPICNN [10] are blurry in occluded regions, such as the leaves in *Herbs* and the metal bin in *Bicycle*. Our method achieves best quantitative and qualitative performance on the synthetic HCI dataset, and shows robustness to occlusions and texture-less surfaces.
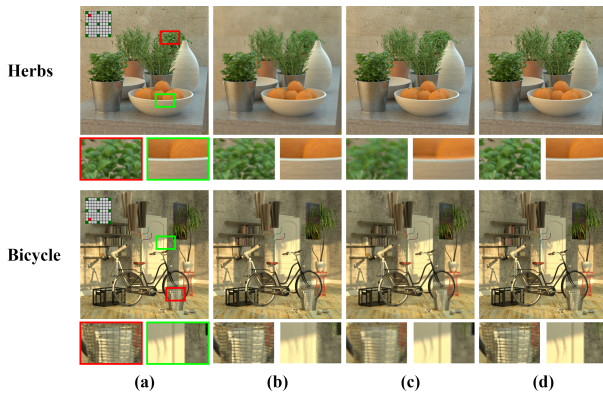
Fig. 6: Visual comparison on the synthetic HCI dataset [31]. (a) Ground-truth. (b) Shearlet [8]. (c) LFEPICNN [10]. (d) Ours

V. Conclusions

In this work, we proposed a novel self-supervised framework to reconstruct dense light fields by synthesizing novel intermediate light field views. To adopt small-sized light field datasets, we introduced the cycle consistency mechanism to fine-tune a pre-trained video interpolation method in a self-supervised fashion. In this context, this method does not require paired ground-truth and is able to use for any low angular resolution light field input. The proposed method outperforms other state-of-the-art approaches on various light fields, in particular, given handling wide disparity inputs. In addition, our method can be adopted to any video interpolation approach, and let any 2D video interpolation into apply to light field data. For future work, we may focus on adopting the proposed method to more challenging scenarios, such as very sparse light fields captured by camera arrays. This may require additional priors handle the sparsity.

References

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

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

[3] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 926–954, 2017.

[4] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light Field Photography with a Hand-Held Plenoptic Camera," Stanford University CSTR, Tech. Rep., Apr. 2005.

[5] T. Herfet, T. Lange, and K. Chelli, "5D light field video capture," in *Proceedings of the 16th ACM SIGGRAPH European Conference on Visual Media Production*, 2019.

[6] M. Alain and A. Smolic, "Light field super-resolution via LFBM5D sparse coding," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2501–2505.

[7] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 24–32.

[8] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 133–147, 2017.

[9] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–10, 2016.

[10] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, "Light field reconstruction using convolutional network on epi and extended applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1681–1694, 2018.

[11] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 333–348.

[12] Z. Cheng, Z. Xiong, C. Chen, and D. Liu, "Light field super-resolution: A benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[14] Y.-L. Liu, Y.-T. Liao, Y.-Y. Lin, and Y.-Y. Chuang, "Deep video frame interpolation using cyclic frame generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8794–8802.

[15] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 892–900.

[16] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous fourier domain," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 1, pp. 1–13, 2014.

[17] H. Wing Fung Yeung, J. Hou, J. Chen, Y. Ying Chung, and X. Chen, "Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 137–152.

[18] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," *arXiv preprint arXiv:1805.09817*, 2018.

[19] 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 Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–14, 2019.

[20] Y. Gao and R. Koch, "Parallax view generation for static scenes using parallax-interpolation adaptive separable convolution," in *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2018, pp. 1–4.

[21] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 670–679.

[22] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 261–270.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1701–1710, 2018.

[25] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9000–9008.

[26] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in *IEEE Conferene on Computer Vision and Pattern Recognition*, 2019.

[27] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Proceedings of the International Conference on Quality of Multimedia Experience*, 2016.

[28] "Inria Lytro Illum dataset," http://www.irisa.fr/temics/demos/lightField/CLIM\/DataSoftware.html, accessed: 26-01-2018.

[29] "The stanford light field archive," http://lightfield.stanford.edu/lfs.html, accessed: 05-03-2019.

[30] P. Matysiak, M. Grogan, M. Le Pendu, M. Alain, and A. Smolic, "A pipeline for lenslet light field quality enhancement," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 639–643.

[31] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Asian Conference on Computer Vision*. Springer, 2016.

[32] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldluecke, "Light field intrinsics with a deep encoder-decoder network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9145–9154.