Depth from Defocus Approaches for Video Depth Estimation

Zhengyang Wei

Abstract—Depth estimation remains challenging to perform with a single image due to the loss of three-dimension information. Inspired by depth from defocus and the scale-consistent video-based learning, we propose a video depth estimation method called DfD-SC-Depth, which applies unsupervised training with monocular snippets on the Depth from Defocus model. We demonstrate the performance of our proposed method and the baseline on the NYUv2 dataset and analyze the results qualitatively and quantitatively. Thanks to the combination of the strength of each technique, our method achieves better performance. Furthermore, this approach offers a possibility for using a phase-coded aperture camera's video to improve the depth from defocus model without extra depth information.

Index Terms—Depth from Defocus, Video Depth Estimation, Video-based Learning

1 Introduction

.

Video depth information is important for robotics, autonomous driving, 3D reconstruction, and beyond [1]. Depth information can be acquired by expensive sensors like LIDAR and stereo cameras. Generating high-quality depth-from-color can inexpensively complement these sensors.

Monocular depth estimation(MDE) from a single color image is challenging. Estimating absolute depth from a color image is ill-posed without a second image for triangulation [2]. Most MDE approaches rely on contextual cues to evaluate the relative location of objects in an image [3]. Depth from defocus processes are used for single image depth estimation and outperform many state-of-art methods [4]. Well-designed end-to-end coded aperture MDE technique can take full advantage of the monocular depth cue [5] and achieve higher accuracy for depth estimation.

Video depth estimation considers the depth of a single frame and the Spatio-temporal relationship between adjacent video frames, which improves the performance of depth estimation [2]. Furthermore, video-based learning doesn't have an impact on single-image-based tasks. Still, it's critical for video-based applications [6], which makes it a practical approach to improve the performance of depth estimation.

However, the depth from defocus method hasn't been applied to video video-based learning for depth estimation yet.

This paper proposes a video depth estimation method called DfD-SC-Depth, which applies unsupervised training with monocular snippets on the Depth from Defocus model and compares the proposed method with our baselines qualitative and quantitatively. Specifically, our contributions are the following:

1) We proposed a framework that combined the unsupervised video-based training based on scale consistency with the depth from defocus method.
2) We conduct comparison experiments on NYUv2 dataset [7] and qualitatively and quantitatively analyze the results.
3) We proposed a feasible plan to improve models with videos captured by the phase-coded aperture camera without depth information.

• *Z. Wei is with the Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, 94305.*
E-mail: zywei@stanford.edu

2 Related Work

2.1 Depth from defocus method for MDE

Besides pictorial cues, defocus blur is a vital depth cue for monocular depth estimation. The camera can produce defocus blur according to the depth of the object and camera settings when capturing images. So, applying it to depth estimation has become a trend.

Using images with defocus blur in deep learning approaches outperforms the use of all-in-focus images [8]. Conventional lenses or hand-crafted phase-coded apertures have been regarded as powerful depth estimation techniques [9]. With the development of deep optics, the End-to-End coded aperture was used to improve the defocus blur and encode more information [10]. Ikoma et al. proposed a framework for single RGB image depth estimation, including an occlusion-aware image formation model, a rotationally symmetric phase-coded aperture, and the corresponding preconditioning approach [5]. It can also estimate an all-in-focus image, which is helpful to extract the Spatio-temporal relationship between adjacent frames. The depth from defocus method is one of the most promising ways to solve the ill-posed problem.

2.2 Video depth estimation

Unlike depth estimation for a single image, video depth estimation usually takes information between frames into account. As a result, researchers attempt to design many

self-supervised or unsupervised approaches to realize the optimization of depth estimation models, mainly utilizing the frame information of videos.

Monodepth2 uses minimum reprojection loss to tackle occlusions between frames and auto-masking loss to ignore stationary pixels [2]. Packnet has symmetrical packing and unpacking blocks, and it uses the neighbor frames s temporal context to realize self-supervised scale-aware structure-from-motion [11]. SC-Depth penalizes the inconsistency of predicted depths of adjacent with frames geometry consistency loss [6]. M4Depth maintains the Spatio-temporal consistency with time recurrence and motion information [12]. Robust CVD jointly optimize camera poses as well as depth deformation in 3D and resolve fine-scale details using a geometry-aware depth filter [13]. However, almost all the video depth estimation methods haven't considered the depth from defocus approaches.

3 Proposed Method

In this section, we describe our proposed DfD-SC-Depth model. The overall pipeline is illustrated in Fig.1.

3.1 Depth from Defocus Model

3.1.1 Image Formation Models

The imaging system's depth-dependent point spread function(PSF) can be controlled by the surface height variation of the diffractive optical element DOE of the phase-coded aperture [14]. [14] proposes to use Equ.1 to model the PSF.

$$PSF(\rho, z, \lambda) = |\frac{2\pi}{\lambda s}\int_0^{+\infty} rD(r,\lambda,z)P(r,\lambda)J_0(2\pi\rho r)dr|^2 \tag{1}$$

ρ and r are the radial distance on the sensor and aperture planes; λ is the wavelength; s is the distance between lens and sensor, and J_0 is the zeroth-order Bessel function. $D(r,\lambda,z)$ is the defocus factor modeling the depth variation of the PSF for points at a distance z from the lens and given by

$$D(r,\lambda,z) = \frac{z}{\lambda(r^2+z^2)} e^{i\frac{2\pi}{\lambda}(\sqrt{r^2+z^2}-sqrtr^2+d^2)} \tag{2}$$

[15] proposed a radially symmetric DOE design which reduces the required memory and time for optimization. The corresponding phase delay $P(r,\lambda)$ is defined as

$$P(r,\lambda) = a(r)e^{i\frac{2\pi}{\lambda}(n(\lambda)-n_{air})h(r)} \tag{3}$$

Here n_{air} is the refractive index of air, and $a(r)$ is the transmissivity of the phase mask. [5] adopts nonlinear differentiable image formation model to the wavelength- and depth-dependent PSF, which is given by Equ.4, where l_k is the sub-images, and α_k is the binary masks composed by quantized depth maps.

$$\begin{aligned} E_k(\lambda) &:= PSF_k * \sum_{k'=0}^{k} \alpha_{k'} \\ \widetilde{l}_k &:= \frac{PSF_k(\lambda) * l_k}{E_k(\lambda)} \\ \widetilde{a}_k &:= \frac{PSF_k(\lambda) * \alpha_k(\lambda)}{E_k(\lambda)} \\ b(\lambda) &= \sum_{k=0}^{K-1} \widetilde{l}_k \prod_{k'=k+1}^{K-1} (1-\widetilde{a}_k') + \eta \end{aligned} \tag{4}$$

The nonlinear model can generate realistic defocus images from color images and their corresponding densely labeled depth maps for our method.

3.1.2 UNet-based Estimation for Image and Depth

We adopt the architecture proposed in [5].For defocus images produced by the nonlinear image formation models, Tikhonov-regularized least squares method is used to map the 2D image into a multiplane representation $l^{(est)} \in \mathbb{R}^{M\times N\times K}$, as shown in Equ.5. It can generate a layered representation with sharper details.

$$l^{(est)} = \underset{l\in\mathbb{R}^{M\times N\times K}}{\arg\min} \quad \|b - \sum_{k=0}^{K-1} PSF_k * l_k\|^2 + \gamma \ \|l\|^2 \tag{5}$$

Then, both the defocus image and its corresponding multiplane representation are concatenated, which is the input for the UNet architecture. And the detail of the implementation of the network is shown in TABLE 1.

3.2 Scale-consistent Depth Learning from Video

3.2.1 Framework overview

The framework is aimed to train the DfD depth network and pose network from unlabeled videos. For two adjacent frames (I_a, I_b) which are randomly sampled from a video, the model can respectively estimate their depth maps(D_a, D_b) and relative camera poseP_{ab}. So the the reference depth D_b^a can be synthesized with the source depth D_a by differentiable warping [6]. Then the reference depth maps can be used for training the whole model as shown in Fig. 1.

3.2.2 Training Loss Function

[6] proposed a kind of objective function for scale-consistent model, which is formulated as follows:

$$L = \alpha L_P^M + \beta L_S + \gamma L_G \tag{6}$$

α, β, γ are the loss weighting terms.

L_P^M stands for the photometric loss L_P weighted by the proposed M_s and is defined as:

$$L_P^M = \frac{1}{|\mathcal{V}|}\sum_{p\in\mathcal{V}} (M_s(p)\cdot L_P(p)) \tag{7}$$

with the synthesized I_a' and the reference image I_a from predicted depth D_a and pose P_{ab}, L_P is formulated as

$$L_P = \frac{1}{|\mathcal{V}|}\sum_{p\in\mathcal{V}} \left(\lambda \left\|I_a(p) - I_a'(p)\right\|_1 + (1-\lambda)\frac{1-\mathrm{SSIM}_{aa'}(p)}{2}\right) \tag{8}$$

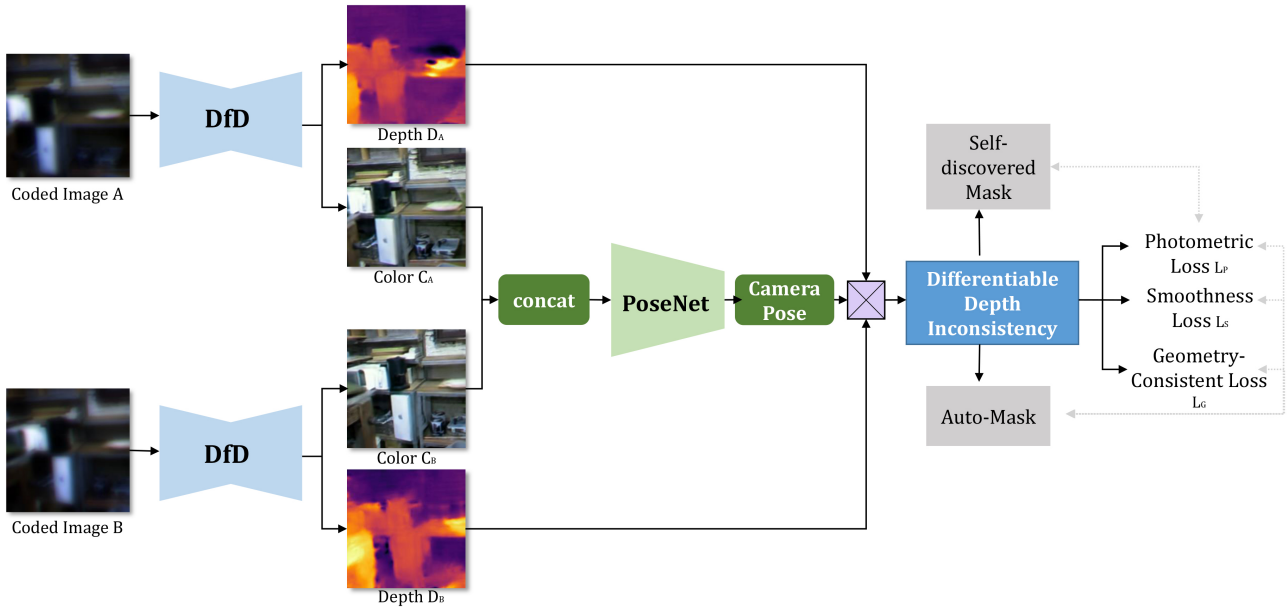

Fig. 1. Our Framework. Given two adjacent coded images randomly sampled from the video, their depth maps and all-in-focus images are first estimated by the DfD models. Then the relative camera pose is calculated by the PoseNet. With the predicted depth and pose, the reference depth for the first frame is synthesized with the estimated depth of the second frame and camera pose. Then the network is supervised by the photometric loss weighted by a self-discover mask, the smoothness loss, and the geometry consistent loss. Finally, the losses are averaged over valid areas determined by an auto-mask.

$$\mathrm{SSIM}(\mathrm{x},\mathrm{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{9}$$

where $\mathcal{V}$ is the set of valid points that are successfully projected from I_a to the image plane of I_b; p is the generic point in $\mathcal{V}$; x, y stands for two 3×3 patches around the central pixel; C_1 and C_2 are constants; μ and σ are mean and variance of the image color respectively. And the depth inconsistency map D_{diff} for each $p \in \mathcal{V}$ and self-discovered mask (M_s) are defined as:

$$D_{\text{diff}}(p) = \frac{|D_b^a(p) - D_b'(p)|}{D_b^a(p) + D_b'(p)} \tag{10}$$

$$M_s = 1 - D_{\text{diff}} \tag{11}$$

Here, D_b^a is the synthesized depth for I_b with D_a and pose P_{ab} by the underlying rigid transformation. D_b' is an aligning interpolation of D_b.

L_S stands for the edge-aware smoothness loss. Formally,

$$L_S = \sum_p \left(e^{-\nabla I_a(p)} \cdot \nabla D_a(p)\right)^2 \tag{12}$$

where ∇ is the first derivative along spatial directions.

L_G is the geometric consistency loss which can be computed with the inconsistency map.

$$L_G = \frac{1}{|\mathcal{V}|} \sum_{p \in \mathcal{V}} D_{\text{diff}}(p) \tag{13}$$

which minimizes the geometric inconsistency of predicted depths for two adjacent frames.

What's more, the loss is averaged over valid points, which are determined by the auto-mask M_a proposed in (Godard et al. 2019). For each $p \in \mathcal{V}$, we have

$$M_a(p) = \begin{cases} 1 & \text{if } \|I_a(p) - I_a'(p)\|_1 < \|I_a(p) - I_b(p)\|_1 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where M_a is a binary mask for each point in $\mathcal{V}$, and I_a' is the warped image from I_b using the estimated depth and pose. It removes the points where the identity mapping results in a lower loss.

4 Experimental Results

This section discusses the design of our experiments and then analyzes and evaluates the qualitative and quantitative results.

4.1 Dataset

The NYU-Depth V2 (NYUv2) data set is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras [7]. We choose it for our experiments since it has densely labeled depth maps which can help us generate defocus images. And we use the officially provided 654 densely labeled images for testing. For the training data, we use videos of 27 different scenes. The validation data is from the rest video and has the size of 10% of the training data. And we fill the depth maps for the dataset with NYU Depth V2's toolbox according to Anat Levin's colorization approach[1].

1. https://www.cs.huji.ac.il/w yweiss/Colorization/

TABLE 1
UNet Achitecture for proposed DfD model (Ignoring the reduce of height and width caused by convolution, and $N = 3 + 3\#$multiplane representation)

Input	Layer	Output
concat(image, inv) $H \times W \times N$	Conv + ReLU $N \times N \times 3 \times 3$	Tensor 1 $H \times W \times N$
Tensor 1 $H \times W \times N$	Conv + ReLU $32 \times N \times 3 \times 3$	Tensor 2 $H \times W \times 32$
Tensor 2 $H \times W \times 32$	(Conv + ReLU)$\times$ 2 $32 \times 32 \times 3 \times 3$	Tensor 3 $H \times W \times 32$——1
Tensor 3 $H \times W \times 32$	MaxPool	Tensor 4 $H/2 \times W/2 \times 32$
Tensor 4 $H/2 \times W/2 \times 32$	Conv + ReLU $64 \times 32 \times 3 \times 3$	Tensor 5 $H/2 \times W/2 \times 64$
Tensor 5 $H/2 \times W/2 \times 64$	Conv + ReLU $64 \times 64 \times 3 \times 3$	Tensor 6 $H/2 \times W/2 \times 64$
Tensor 6 $H/2 \times W/2 \times 64$	MaxPool	Tensor 7 $H/4 \times W/4 \times 64$
Tensor 7 $H/4 \times W/4 \times 64$	(Conv + ReLU)$\times$ 2 $64 \times 64 \times 3 \times 3$	Tensor 8 $H/4 \times W/4 \times 64$
Tensor 8 $H/4 \times W/4 \times 64$	MaxPool	Tensor 9 $H/8 \times W/8 \times 64$
Tensor 9 $H/8 \times W/8 \times 64$	Conv + ReLU $128 \times 64 \times 3 \times 3$	Tensor 10 $H/8 \times W/8 \times 128$
Tensor 10 $H/8 \times W/8 \times 128$	Conv + ReLU $128 \times 128 \times 3 \times 3$	Tensor 11 $H/8 \times W/8 \times 128$
Tensor 11 $H/8 \times W/8 \times 128$	MaxPool	Tensor 12 $H/16 \times W/16 \times 128$
Tensor 12 $H/16 \times W/16 \times 128$	(Conv + ReLU) $\times$ 2 $128 \times 128 \times 3 \times 3$	Tensor 13 $H/16 \times W/16 \times 128$
Tensor 13 $H/16 \times W/16 \times 128$	UpSample Bilinear	Tensor 14 $H/8 \times W/8 \times 128$
concat(Tensor 11,14) $H/8 \times W/8 \times 256$	Conv + ReLU $128 \times 256 \times 3 \times 3$	Tensor 15 $H/8 \times W/8 \times 128$
Tensor 15 $H/8 \times W/8 \times 128$	Conv + ReLU $128 \times 128 \times 3 \times 3$	Tensor 16 $H/8 \times W/8 \times 128$
Tensor 16 $H/8 \times W/8 \times 128$	UpSample Bilinear	Tensor 17 $H/4 \times W/4 \times 128$
concat(Tensor 17,8) $H/4 \times W/4 \times 192$	Conv + ReLU $64 \times 192 \times 3 \times 3$	Tensor 18 $H/4 \times W/4 \times 64$
Tensor 18 $H/4 \times W/4 \times 64$	Conv + ReLU $64 \times 64 \times 3 \times 3$	Tensor 19 $H/4 \times W/4 \times 64$
Tensor 19 $H/4 \times W/4 \times 64$	UpSample Bilinear	Tensor 20 $H/2 \times W/2 \times 64$
concat(Tensor 20,6) $H/2 \times W/2 \times 128$	Conv + ReLU $64 \times 128 \times 3 \times 3$	Tensor 21 $H/2 \times W/2 \times 64$
Tensor 21 $H/2 \times W/2 \times 64$	Conv + ReLU $64 \times 64 \times 3 \times 3$	Tensor 22 $H/2 \times W/2 \times 64$
Tensor 22 $H/2 \times W/2 \times 64$	UpSample Bilinear	Tensor 23 $H \times W \times 64$
concat(Tensor 23,3) $H \times W \times 96$	Conv + ReLU $32 \times 96 \times 3 \times 3$	Tensor 24 $H \times W \times 32$
Tensor 24 $H \times W \times 32$	Conv + ReLU $32 \times 32 \times 3 \times 3$	Tensor 25 $H \times W \times 32$
Tensor 25 $H \times W \times 32$	Conv + ReLU $4 \times 32 \times 3 \times 3$	Tensor 26 $H \times W \times 4$

4.2 Baseline Comparisons

We compare our method to the original methods:

1) Ikoma et al [5]. A framework for single RGB image depth estimation, including an occlusion-aware image formation model, a rotationally symmetric phase-coded aperture, and the corresponding preconditioning approach. We call this method DfD for convenience.
2) SC-Depth [6]. An unsupervised method with video data penalizes the inconsistency of predicted depths of adjacent with frames geometry consistency loss.

4.3 Metrics

We choose RMSE, AbsRel, Log10 and a_1, a_2, a_3 as the metrics.

1) RMSE is a quadratic scoring rule that measures the error's average magnitude and is defined as follow:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)} \tag{15}$$

2) AbsRel determines how bad the error is and doesn't depend on the size of the quantity. It's defined as

$$AbsRel = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{\hat{y}_i} \tag{16}$$

3) Log10 only cares about the percentual difference and is defined as

$$Log10 = \frac{1}{N}\sum_{i=1}^{N}|\log_{10} y_i - \log_{10} \hat{y}_i| \tag{17}$$

4) a_1, a_2, a_3

$$\begin{aligned} Thresh_i &= \max\left(\frac{y_i}{\hat{y}_i}, \frac{\hat{y}_i}{y_i}\right) \\ a_j &= \frac{\#Thresh_i < 1.25^j}{N}, j = 1, 2, 3 \end{aligned} \tag{18}$$

4.4 Qualitative Results

4.4.1 Results

The qualitative comparison with the baselines is shown in Fig. 2 and the detail of the comparison is shown in Fig. 3.

4.4.2 Analysis and Evaluation

Our DfD-SC-Depth method visually obtains the depth maps closest to the ground truth. DfD-SC-Depth can protect the sharper edges for the estimated depth maps and get more accurate depth for the area far from the camera compared with SC-Depth. DfD-SC-Depth can also modify some apparent errors in the depth maps of the DfD model.

Our method can successfully capture edges of the doors, walls, windows in Row 1, 4, 5 of Fig. 2, and the details are shown in Fig. 3; while SC-Depth leads to blurring artifact and DfD method leads to aliasing artifact.

Our DfD-SC-Depth can also preserve the shape of objects in the images. For example, we can tell from the chairs and tables obtained by our method easily in Row 1, 6 of Fig. 2. However, it's difficult to point out where the chairs are in depth-maps obtained by DfD and SC-Depth in Row 1 of Fig. 2.

SC-Depth sometimes might fail to accurately get the depth for the area far from the camera. For example, in Row 2, 3, 4 of Fig. 2, the deepest regions of the SC-Depth's depth map are much different from the ground truth, but our method's results are almost consistent with the ground truth in these regions.

There are some apparent errors in the depth map of the DfD method, such as the right-bottom of Row 2 of Fig. 2 and right-top of Row 7 of Fig. 2. Fortunately, our DfD-SC-Depth can also effectively modify these artifacts.

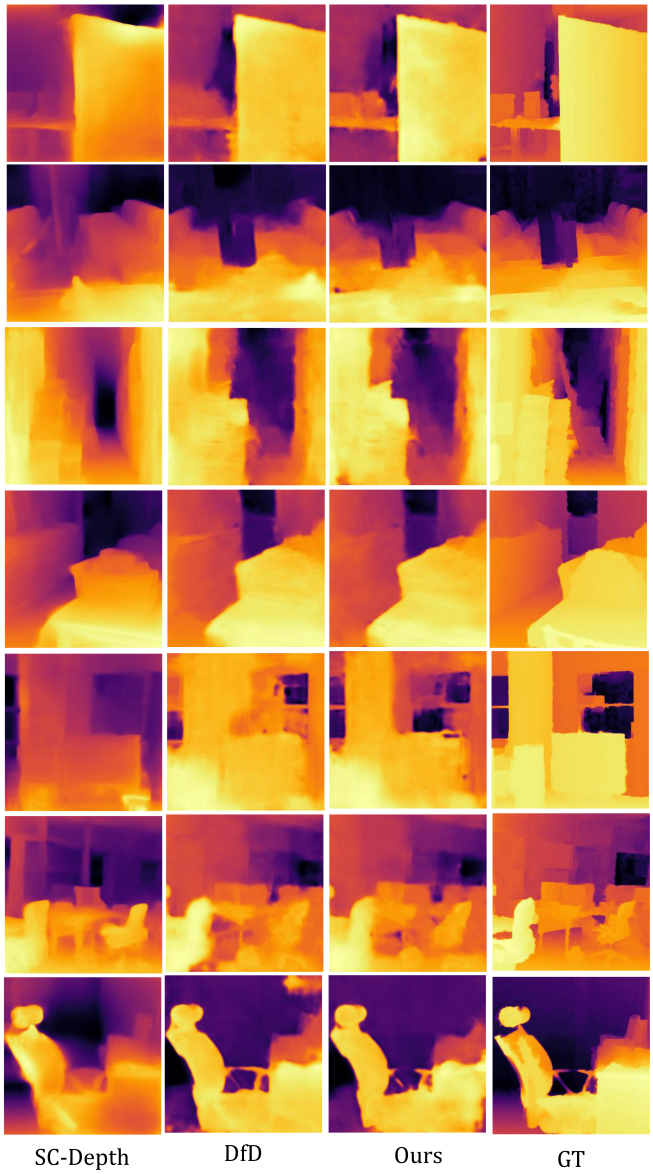

Fig. 2. Qualitative comparison with baselines

4.5 Quantitative Results

4.5.1 Results

The quantitative comparison with baselines is shown in TABLE 2.

TABLE 2
Quantitative comparison with baselines

	SC-Depth	DfD	DfD-SC-Depth
$RMSE \downarrow$	0.399	0.426	**0.372**
$AbsRel \downarrow$	0.214	0.215	**0.198**
$Log10 \downarrow$	**0.089**	0.102	0.095
$a_1 \uparrow$	0.682	0.649	**0.688**
$a_2 \uparrow$	0.828	0.830	**0.846**
$a_3 \uparrow$	**0.927**	0.916	0.923

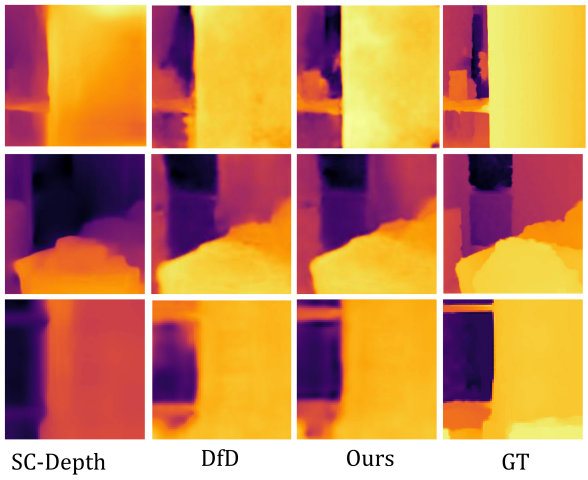

Fig. 3. Qualitative comparison with baselines

4.5.2 Analysis and Evaluation

As shown in TABLE 2, the RMSE, AbsRel, a_1, and a_2 of our proposed DfD-SC-Depth method are much better than the baselines. However, Log10 and a_3 of SC-Depth are a little better than our method, which might be caused by the fact that SC-Depth can yield fewer points that deviate too far from the ground truth. However, our approach can achieve better accuracy overall.

5 Conclusion

5.1 Limitations and Future Work

First, this method is sensitive to the learning rate since there is no ground truth for training data, and the reference depth is generated by the predicted depth and pose. So if the learning rate is too large, the model might jump in the wrong direction.

Second, we use the coded defocus images generated by RGBD images. And since the size of densely labeled images we can acquire is very limited, our experiment isn't conducted on a significant number of unlabeled training data as expected. Therefore, the final results may not reach the best situation.

In the future, if we can use the phase-coded aperture camera to capture videos, we will validate our method with more training data. We also consider combining other types of video consistency information with depth from defocus methods.

5.2 Conclusion

We successfully applied unsupervised video-based training on depth from the defocus method. We experimentally demonstrate that our approach produces better performances than the baseline on the NYUv2 dataset. Even though the improvement is relatively slight, the method is still inspiring. Using the phase-coded aperture camera to capture videos or continuous images, we might improve the models without extra depth information. That ensures the method is feasible for real-life applications and can be widely applied to specific scenes.

ACKNOWLEDGMENTS

I would like to express my greatest gratitude to Prof. Gordon Wetzstein for his fascinating lectures and well-prepared projects and my project mentor, Mark Nishimura, for giving me useful insights.

REFERENCES

[1] A. Sabnis and L. Vachhani, "Single image based depth estimation for robotic applications," in *2011 IEEE Recent Advances in Intelligent Computational Systems*, 2011, pp. 102–106.

[2] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth prediction," October 2019.

[3] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *CoRR*, vol. abs/1406.2283, 2014. [Online]. Available: http://arxiv.org/abs/1406.2283

[4] P. Trouvé, F. Champagnat, G. L. Besnerais, J. Sabater, T. Avignon, and J. Idier, "Passive depth estimation using chromatic aberration and a depth from defocus approach." *Applied optics*, vol. 52 29, pp. 7152–64, 2013.

[5] H. Ikoma, C. M. Nguyen, C. A. Metzler, Y. Peng, and G. Wetzstein, "Depth from defocus with learned optics for imaging and occlusion-aware depth estimation," *IEEE International Conference on Computational Photography (ICCP)*, 2021.

[6] J.-W. Bian, H. Zhan, N. Wang, Z. Li, L. Zhang, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth learning from video," *International Journal of Computer Vision (IJCV)*, 2021.

[7] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[8] M. Carvalho, B. L. Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat, "Deep depth from defocus: how can defocus blur improve 3d estimation using dense neural networks?" 2018.

[9] A. Levin, S. W. Hasinoff, P. Green, F. Durand, and W. T. Freeman, "4d frequency analysis of computational cameras for depth of field extension," *ACM Trans. Graph.*, vol. 28, no. 3, jul 2009. [Online]. Available: https://doi.org/10.1145/1531326.1531403

[10] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "Phasecam3d — learning phase masks for passive single view depth estimation," in *2019 IEEE International Conference on Computational Photography (ICCP)*, 2019, pp. 1–12.

[11] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[12] M. Fonder, D. Ernst, and M. V. Droogenbroeck, "M4depth: A motion-based approach for monocular depth estimation on video sequences," May 2021.

[13] J. Kopf, X. Rong, and J.-B. Huang, "Robust consistent video depth estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[14] J. W. Goodman, "Introduction to fourier optics," *Introduction to Fourier optics, 3rd ed., by JW Goodman. Englewood, CO: Roberts & Co. Publishers, 2005*, vol. 1, 2005.

[15] X. Dun, H. Ikoma, G. Wetzstein, Z. Wang, X. Cheng, and Y. Peng, "Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging," *Optica*, vol. 7, no. 8, pp. 913–922, Aug 2020. [Online]. Available: http://opg.optica.org/optica/abstract.cfm?URI=optica-7-8-913