学習ベース両眼ステレオが持つ事前知識の NeRF への導入

藤村 友貴[1,a)]　櫛田 貴弘[2]　北野 和哉[1]　舩冨 卓哉[1]　向川 康博[1]

概要

本研究では学習ベース両眼ステレオがもつ事前知識を Neural Radiance Field (NeRF) の学習に利用する手法を提案する．一般的な NeRF はシーンごとに学習を行うが，これに対し大規模なデータセットで学習された単眼深度推定などのモデルをシーンの事前知識として利用する手法が近年提案されている．本研究ではシーンの新たな事前知識として，大規模なデータセットで学習された両眼ステレオの利用を試みる．両眼ステレオに NeRF で合成したステレオペアを入力し，推定した視差を用いて新たな学習画像を生成する．本手法を既存手法に適用することで，入力画像の枚数が少ない場合における新規視点合成の精度が向上することを示す．

1. はじめに

Neural Radiance Field (NeRF) [6] とは，多視点で撮影された画像を入力として，ニューラルネットワークで表現された輝度と密度の場を求めることで，任意視点での画像の生成を可能とする技術である．NeRF の課題の一つとして，入力画像の枚数が少ない場合は大きく精度が低下してしまうことが知られている．

この問題に対し，大規模なデータセットで学習されたモデルの事前知識を利用する研究が行われている．通常の NeRF はシーンごとに学習を行うが，データセットで学習された事前知識を導入することで，入力が不足する問題に対処する．例えば，事前学習された単眼深度推定 [7] の出力を利用する研究が行われており [3], [9], [10], [11]，推定した深度を幾何的な制約として利用する．

これらに対し本研究は，学習ベース両眼ステレオが持つ事前知識を NeRF の学習に利用しようとする新たな試みである．両眼ステレオとは，2 枚のステレオ画像のペアから視差を推定する手法である．本研究では，学習後の NeRF が生成したステレオペアに対して学習ベースの両眼ステレオ [5] を適用する．入力画像の枚数が少ない場合は NeRF が生成する画像にはノイズが含まれるが，このようなノイズに対して学習ベース両眼ステレオは頑健に視差を推定することができる（図 1）．本研究ではこの性質を実験的に明らかにし，推定された視差を用いて学習に使用した視点の画像を変形し，新たな学習画像として利用する手法を提案する．

[1] 奈良先端科学技術大学院大学
[2] 立命館大学
[a)] fujimura.yuki@is.naist.jp

2. 提案手法

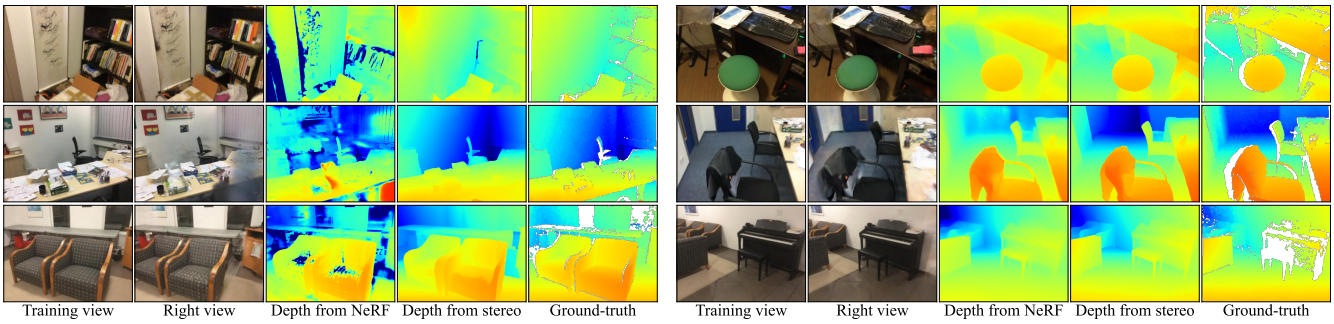
2.1 学習ベース両眼ステレオの NeRF 生成画像への適用

両眼ステレオは平行化された 2 枚の画像を入力に必要とする．そこで本研究では，学習後の NeRF を用いて，学習に使用した視点とその視点から水平方向に微小にずらした視点からの画像を生成し，このステレオペアを学習済みの両眼ステレオに入力する．

図 1 に学習ベース両眼ステレオの一つである RAFT-Stereo [5] を適用した例を示す．表 1 には定量評価を示す．ここでは，推定した視差を既知の焦点距離とカメラ間の距離（基線長）を用いて深度に変換している．左から，学習に用いた視点での再構成画像，そこから右に微小にずらした視点での生成画像，NeRF でレンダリングした深度画像，最初の 2 枚に対して RAFT-Stereo を適用して得られた深度画像，正解の深度画像である．NeRF の既存手法である (a) K-planes [2] と (b) DäRF [9] に対して，ScanNet [1] の 3 つのシーンで実験を行った．各シーンの学習画像の枚数は 18 枚から 20 枚であり，NeRF の学習としては入力画像の枚数が少ない．これらの結果が示すように，ステレオ画像にノイズが含まれる場合でも，RAFT-Stereo を用いることでより高い精度で深度の推定が可能である．本研究ではこの性質を用いて，NeRF のさらなる精度向上を試みる．

2.2 手法の概要

単純なアプローチとして，ステレオで推定した深度画像を幾何的な拘束として加えることが考えられるが，あとで述べるようにこの方法では学習に加えた深度画像に NeRF が過学習してしまうことがわかった．そこで，幾何的な拘束として用いるのではなく，図 2 に示すように学習視点で再構成した画像を視差で変形し，新たな学習画像として NeRF の再学習を行う手法を提案する．具体的には以下のステップで NeRF の学習を行う．(1) 既存の NeRF のモデ



(a) K-planes [2]　　(b) DäRF [9]

図 1　ScanNet [1] で学習した (a) K-planes [2] と (b) DäRF [9] について，学習後に生成したステレオペアを RAFT-Stereo [5] に入力した例.

表 1　K-planes [2] と DäRF [9] について，ScanNet [1] の学習後にレンダリングした深度と，ステレオ画像を生成し RAFT-Stereo [5] を適用して推定した深度の定量評価

	AbsRel ↓	SqRel ↓	RMSE ↓	RMSE log ↓
K-planes [2]	0.410	1.172	1.449	0.520
K-planes [2] + stereo [5]	0.210	0.375	0.714	0.291
DäRF [9]	0.082	0.029	0.253	0.105
DäRF [9] + stereo [5]	**0.071**	**0.028**	**0.235**	**0.094**

ルの学習を行う．(2) 学習に用いた各視点で画像を再構成する．さらに，視点位置を同じ基線長で左右にずらし 2 枚の画像を生成する．(3) 学習視点で再構成した画像と左右に視点位置をずらして生成した画像をそれぞれペアとして学習ベース両眼ステレオに入力し，2 枚の視差画像を生成する．また，2 枚の視差画像から推定した視差の確信度を計算する．(4) 学習視点の画像と確信度を，推定した視差で順方向に変形し，それらを学習データに加え再度 NeRF の学習を行う．

2.3　視差を用いた新たな学習画像の生成

最初に既存の NeRF の学習後，学習に使用した各視点で画像を再構成する．ここで，各視点で再構成した画像を I_c とする．その後，各視点に対し，ある基線長だけ右にずらした視点から画像 I_r を生成する．各視点におけるペア (I_c, I_r) を学習ベース両眼ステレオに入力し，視差 d_r を推定する．この視差を用いて I_c を順方向に変形する．

$$\hat{I}_r(x + \lfloor d_r(x,y) + 0.5 \rfloor, y) = I_c(x,y) \qquad (1)$$

ここで，$\hat{I}_r$ は変形後の画像であり，(x, y) は画像のピクセル位置である．本研究ではこのようにして得られた変形画像 $\hat{I}_r$ を新たな学習データとして用いる．

ここで，変形した画像を新たな学習画像として用いる有効性について議論する．図 3 に I_c，I_r，$\hat{I}_r$ の例を示す．(a) I_c は学習視点で再構成した画像であり，この視点は NeRF の学習に使用しているため，ノイズの少ない画像が再構成される．(b) I_r は視点を右にずらした視点で生成した画像であり，学習データが少ない場合は，この例のように僅かに視点をずらすだけで，生成される画像には大きなノイズが含まれる．これは，物体が存在しない空間で密度と輝度が値を持ってしまい，雲のようなノイズや色の劣化が生じてしまうためである [10]．一方で，(c) $\hat{I}_r$ は I_c を変形して得られた画像であるため，I_r と比べてノイズが少ない．したがって，$\hat{I}_r$ を学習に用いることで，I_r に含まれていたようなノイズを軽減できると考えられる．なお，本手法は本質的には物体表面の局所的な拡散反射を仮定したものであるが，あとで示すように，このような仮定においても新規視点合成の精度が向上することが実験で確認できた．

2.4　3 視点の一貫性による確信度の計算

両眼ステレオは画像間で遮蔽が生じている箇所は本質的に推定が不可能であり，学習ベース両眼ステレオにおいても精度が低下する．本研究ではこのような遮蔽と両眼ステレオの推定誤差そのものに対処するため，3 視点の一貫性から確信度を計算する．具体的には，視点を右にずらした画像に加え，左にも同じ基線長でずらした画像 I_l を生成する．その後，ステレオペア (I_c, I_l) に対しても両眼ステレオを適用し視差 d_l を推定する．I_r と I_l は同じ基線長で生成した画像であるので，視差 d_r と d_l の間には $d_r(x,y) = -d_l(x,y)$ が成り立つ．この関係を用いてピクセル (x, y) における確信度 $C_c(x, y)$ を以下で計算する．

$$C_c(x,y) = \exp\left(-|d_r(x,y) + d_l(x,y)|\right) \qquad (2)$$

図 2 に示すように，I_c，C_c について式 (1) と同様に左右それぞれで順方向に変形を行い，$\hat{I}_r$ に加えて $\hat{I}_l$，$\hat{C}_r$，$\hat{C}_l$ を生成する．

2.5　誤差関数

元の誤差関数に，生成した画像と確信度を用いた以下の誤差関数を加えて NeRF の再学習を行う．

$$\mathcal{L}_{\text{stereo}} = \frac{1}{2} \sum_{x,y} \Big(\hat{C}_r(x,y)(\hat{I}_r(x,y) - I_r^*(x,y))^2 + \hat{C}_l(x,y)(\hat{I}_l(x,y) - I_l^*(x,y))^2 \Big) \qquad (3)$$

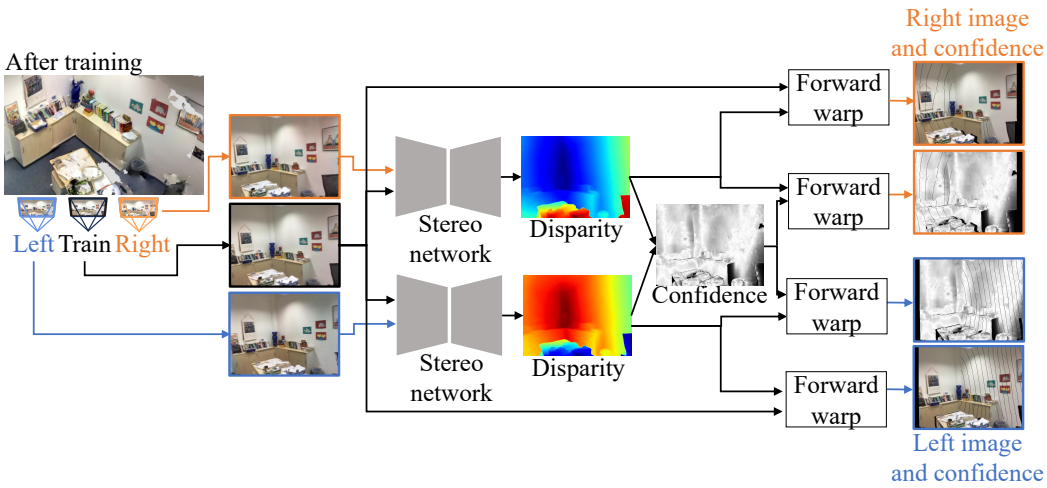
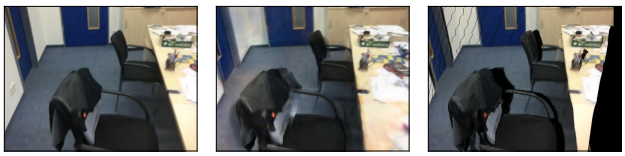


図 2 NeRF が生成したステレオ画像に対する学習ベース両眼ステレオによる視差推定と，推定した視差を用いた新たな NeRF 学習画像の生成.



(a) I_c (b) I_r (c) $\hat{I}_r$

図 3 (a) 学習視点で再構成した画像，(b) 右にずらした視点で生成した画像，(c) 視差による (a) の画像の (b) の視点への変形.

表 2 ScanNet [1] と Tanks and Temples [4] データセットにおける新規視点合成の定量評価．提案手法を適用したものは K-planes [2] + stereo と DäRF [9] + stereo.

	ScanNet [1]		Tanks ans Temples [4]	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Vanilla NeRF [6]	19.03	0.670	17.19	0.559
DDP-NeRF [8]	19.29	0.695	19.18	0.651
SCADE [10]	<u>21.54</u>	0.732	<u>20.13</u>	0.662
K-planes [2]	18.80	0.715	17.27	0.600
K-planes [2] + stereo	19.81	0.738	19.20	0.656
DäRF [9]	21.37	<u>0.764</u>	19.87	<u>0.673</u>
DäRF [9] + stereo	**22.08**	**0.777**	**20.23**	**0.690**

ここで，I_r^*，I_l^* は NeRF が学習中に生成した画像である.

3. 実験

3.1 実装とデータセット

本研究では K-planes [2] と DäRF [9] に提案手法を適用した．学習ベース両眼ステレオについては RAFT-Stereo [5] の学習済みモデルを用いた.

室内のシーンで撮影された二つのデータセット（ScanNet [1]，Tanks and Temples [4]）を用いて実験を行なった．それぞれ DDP-NeRF [8]，SCADE [10] で用いられた 3 つのシーンを用いた．各シーンは学習視点数が 20 前後であり，NeRF の学習としては入力画像の枚数が少ないという問題設定である.

表 3 両眼ステレオで推定した深度を誤差に用いた場合との比較

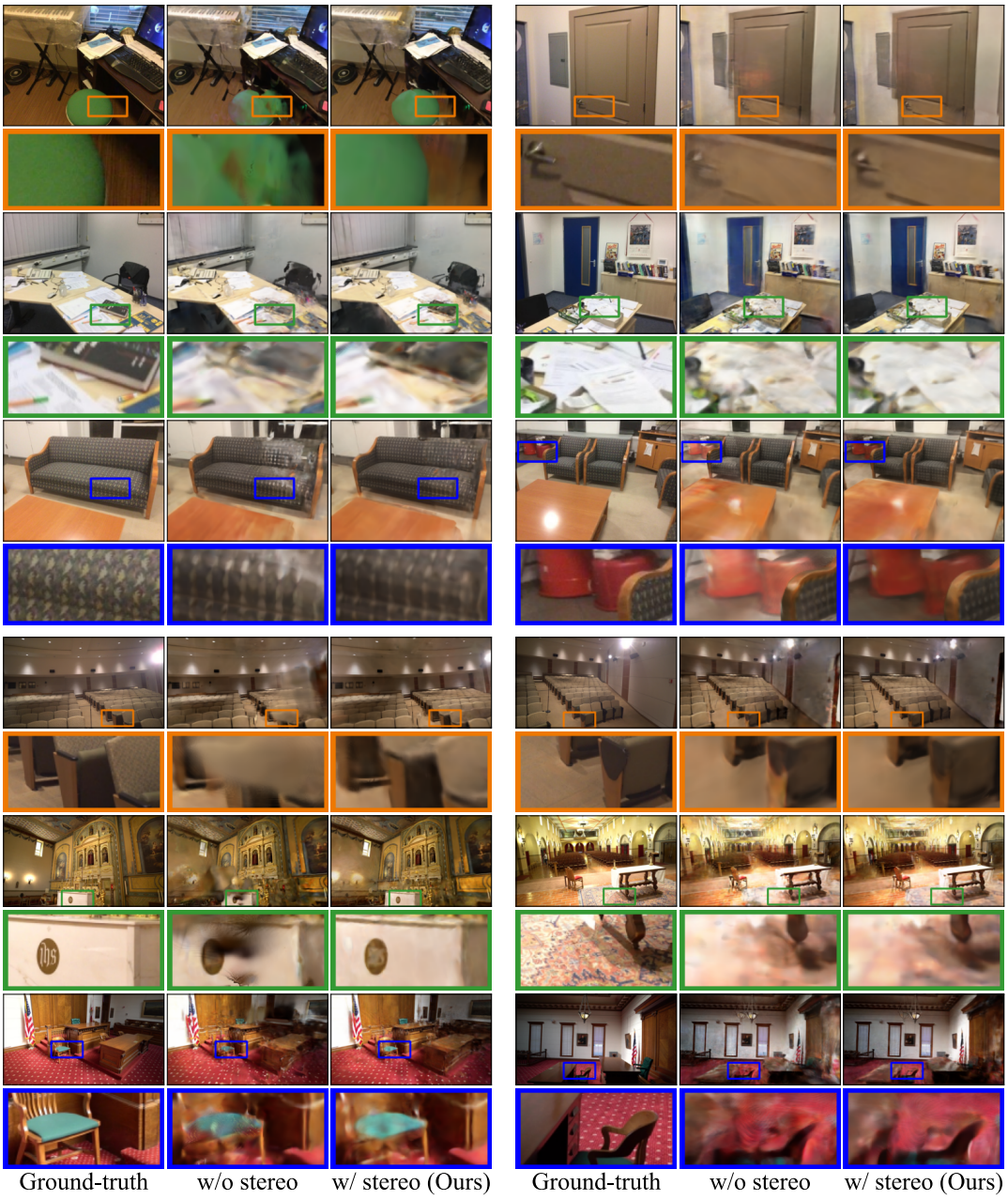
	Novel view synthesis		Depth (test)	Depth (train)
	PSNR ↑	SSIM ↑	RMSE log ↓	RMSE log ↓
w/ depth	21.42	0.762	0.111	0.093
w/o depth	22.08	0.777	0.102	0.097

3.2 実験結果

新規視点合成 表 2 に新規視点合成の実験結果を示す．提案手法を K-planes と DäRF に適用したものがそれぞれ K-planes + stereo，DäRF + stereo である．K-planes のようなシンプルなモデルに対して提案手法を用いた場合，大幅に精度が向上することが確認できる．DäRF は入力が少数という問題設定に対して，単眼深度推定の結果と単眼深度推定そのものを同時に最適化するという複雑なモデルであるが，提案手法によるシンプルな拡張で，さらなる精度向上が可能であることが確認できる．図 4 に (a) K-planes と (b) DäRF を用いて生成した新規視点の画像の例を示す．2.3 で述べたように，提案手法によって雲のようなノイズや色の劣化が低減できることが確認できる.

確信度の影響 提案手法は式 (3) で示したように左右方向で生成した画像，および確信度を誤差関数に用いている．図 5 に学習後に生成された画像の確信度の有無による比較を示す．左右の一貫性を確信度として用いることで，隠蔽により両眼ステレオの精度が低下する深度が不連続であるような箇所において誤差を低減できる.

深度を誤差関数に加えた学習 表 3 に，両眼ステレオで得られた深度を誤差関数に用いた場合との比較を示す．新規視点合成と NeRF がレンダリングした深度についての評価を行なった．この表に示すように，深度を誤差に用いると新規視点合成の精度の低下がみられる．また，テスト視点でレンダリングした深度についても精度が大きく低下している．このことから，視差から計算した深度を直接学習に



(a) K-planes [2]　　(b) DäRF [9]

図 4　ScanNet [1] と Tanks and Temples [4] データセットにおける新規視点合成の定性評価. 提案手法を適用したものは w/ stereo (Ours) で示してある.

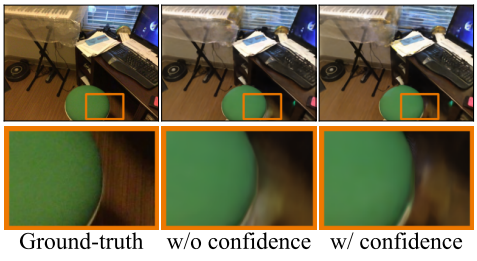


図 5　確信度有り無しでの比較

用いると，学習した視点での過学習が生じてしまうことが推察される.

4. まとめ

本研究では，学習ベース両眼ステレオが持つ事前知識を NeRF の学習に導入する手法を提案した．学習後の NeRF が生成したステレオペアに対して学習ベースの両眼ステレオを適用することで，生成画像に含まれるノイズに対して頑健に視差が推定できる．推定した視差で学習視点の画像を変形し，新たな学習画像として再学習を行うことで，入力画像の枚数が少ない場合における新規視点合成の精度が向上することを示した．本手法は従来手法とは異なる事前知識の導入であり，新たな研究の方向性として期待される.

参考文献

[1] Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T. and Niessner, M.: ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes, *CVPR* (2017).

[2] Fridovich-Keil, S., Meanti, G., Warburg, F. R., Recht, B. and Kanazawa, A.: K-Planes: Explicit Radiance Fields in Space, Time, and Appearance, *CVPR*, pp. 12479–12488 (2023).

[3] Guangcong, Chen, Z., Loy, C. C. and Liu, Z.: SparseNeRF: Distilling Depth Ranking for Few-shot Novel View Synthesis, *ICCV* (2023).

[4] Knapitsch, A., Park, J., Zhou, Q.-Y. and Koltun, V.: Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction, *ACM TOG*, Vol. 36, No. 4 (2017).

[5] Lipson, L., Teed, Z. and Deng, J.: RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching, *3DV* (2021).

[6] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R. and Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, *ECCV* (2020).

[7] Ranftl, R., Bochkovskiy, A. and Koltun, V.: Vision Transformers for Dense Prediction, *ICCV* (2021).

[8] Roessle, B., Barron, J. T., Mildenhall, B., Srinivasan, P. P. and Nießner, M.: Dense Depth Priors for Neural Radiance Fields From Sparse Input Views, *CVPR*, pp. 12892–12901 (2022).

[9] Song, J., Park, S., An, H., Cho, S., Kwak, M.-S., Cho, S. and Kim, S.: DäRF: Boosting Radiance Fields from Sparse Inputs with Monocular Depth Adaptation, *NeurIPS* (2023).

[10] Uy, M. A., Martin-Brualla, R., Guibas, L. and Li, K.: SCADE: NeRFs from Space Carving with Ambiguity-Aware Depth Estimates, *CVPR* (2023).

[11] Yu, Z., Peng, S., Niemeyer, M., Sattler, T. and Geiger, A.: MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction, *NeurIPS* (2022).