Vol.2024-CVIM-237 No.24
2024/3/3

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を利用した Neural Radiance Field (NeRF) の精度向上手法を提案する．多視点画像から新規視点画像を合成する NeRF は，入力画像の枚数が少ない場合に精度が低下する．本研究ではこれに対し，大規模なデータセットで学習された両眼ステレオを利用する手法を提案する．学習ベース両眼ステレオに NeRF で合成したステレオペアを入力して視差を推定する．推定した視差を用いて画像を変形し，これを新たな学習画像として利用する．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．

キーワード：NeRF, learning-based stereo, few-shot

1. はじめに

Neural Radiance Field (NeRF) [13] 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．近年コンピュータビジョン分野で多くの研究がなされており，十分な多視点画像が与えられた場合は高品質な画像を生成することができる．一方で，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 [16] の出力を利用する研究が行われている [19], [22], [25], [32]．単眼深度推定 [4], [15], [16] とは一枚の画像からシーンの深度を推定する手法であり，推定された深度はスケールとシフトの不定性が存在するものの，NeRF の最適化に幾何的な制約として利用することができる．

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 [7], [10], [12], [29] 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た上で，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．

本研究の貢献は以下のとおりである．

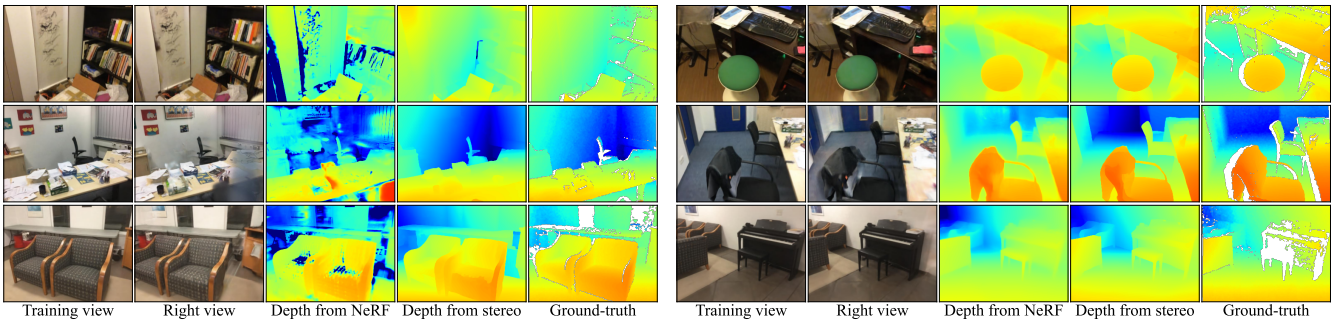
- 学習ベース両眼ステレオが持つ事前知識を NeRF の学習に導入する最初の試みである．
- NeRF が生成したステレオ画像を学習ベースの両眼ステレオに入力した場合，NeRF の生成画像に含まれるノイズに対して頑健に視差が推定できることを実験で明らかにする．
- 推定した視差を用いて学習に使用した視点での生成画像を変形し，新たな学習画像として用いる手法を提案する．
- 入力画像が少ないという問題設定において，既存の NeRF モデルに対する本手法の適用可能性を示し，新規視点合成の精度が向上することを示す．

2. 関連研究

2.1 学習した事前知識の NeRF への導入

入力画像の枚数が少ない場合において NeRF の学習を試みる研究が行われている．主に 2 つのアプローチがあり，1 つは生成される画像や深度画像に対する正則化を導入する手法 [3], [8], [14], [21]，もう 1 つはデータセットで学習されたモデルの事前知識を導入する手

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[2] 立命館大学
Ritsumeikan University
[a)] fujimura.yuki@is.naist.jp



(a) K-planes [5]　　(b) DäRF [19]

図 1 ScanNet [2] で学習した (a) K-planes [5] と (b) DäRF [19] について，学習後に生成したステレオペアを RAFT-Stereo [10] に入力した例．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．

法 [6], [11], [17], [19], [25], [31], [32] である．なお，幾何的な制約や事前知識，及び誤差関数の定式化については Wang ら [24] にまとめられている．

学習された事前知識を用いたアプローチとして近年多く提案されているのが，単眼深度推定 [4], [15], [16] を用いたものである．MonoSDF [32] は多視点画像からのニューラル陰関数表面 [26], [30] の学習において，単眼深度推定の出力結果を誤差関数に導入した．スケールとシフトの不定性はパラメータの更新ごとに最小二乗法によって直接取り除くというアプローチを採用した．SCADE [22] は単眼深度推定の結果を確率的に複数出力するように変更し，単眼深度推定の不定性と透明な物体などで生じる深度の不定性を考慮するモデルを提案した．SparseNeRF [25] は単眼深度推定の出力結果を用いて，深度の順序関係を誤差関数に利用する手法を提案した．DäRF [19] は NeRF の学習と同時に単眼深度推定モデルのさらなる最適化を行い，単眼深度推定モデルがシーンごとのスケールとシフトを学習することを実現した．

これらの手法では単眼深度推定が持つ事前知識をどのように利用するかについての研究がされてきた．一方で本研究は学習ベースの両眼ステレオが持つ事前知識を NeRF の学習に導入しようとする新たな試みである．

2.2 学習ベース両眼ステレオ

2 枚のステレオ画像から視差を推定する両眼ステレオはコンピュータビジョン分野で長らく研究されている．両眼ステレオはステレオマッチングとフィルタリングからなり，ステレオマッチングでは各画像から抽出した特徴量のマッチングを行い，フィルタリングでは密な視差が計算される．学習ベースの両眼ステレオではこれらを end-to-end で実装し，大規模なデータセットで学習することによって精度の向上を図る．従来提案されてきた多くの手法では，ネットワークが出力した中間特徴量を用いてコストボリュームを計算し，2 次元畳み込み [12] や 3 次元畳み込み [1], [7], [33] を用いてコストボリュームをフィルタリングする．したがって，マッチングのための特徴抽出部とフィルタリング部が同時に学習される．これらに対し RAFT-Stereo [10] は特徴量の相関を用いて視差そのものを反復的に更新する設計であり，2 次元畳み込みのみで構成された軽量なモデルである．近年は Transformer [23] を用いたモデルも提案されており，Xu ら [29] は Attention をステレオ画像間にも導入し，畳み込みベースの従来手法とは異なり画像間の関係も考慮した特徴量の学習を実現した．

本研究ではこのような学習ベース両眼ステレオが学習した事前知識を NeRF の学習に導入することを試みる．NeRF と学習ベース両眼ステレオを組み合わせた関連研究として，NeRF を用いて学習用データセットを生成するという研究がある [20]．多視点画像で学習した NeRF は任意視点からの画像とその視点に対応する深度画像を生成可能であるため，学習した NeRF を用いて大量のステレオ画像と深度画像を生成し，両眼ステレオの学習に利用する．本研究ではこれに対し，NeRF が生成したステレオ画像を NeRF の学習そのものに利用する試みである．

3. 提案手法

本研究は，学習ベース両眼ステレオがもつ事前知識を NeRF の学習に導入するという全く新しい手法を提案する．入力画像が少ないという問題設定において，学習ベース両眼ステレオが持つ事前知識を利用する．本章では，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両眼ステレオを適用する．学習ベース両眼ステレオが NeRF 生成画像に含まれるノイズに対して頑健に視差の推定が可能であることを示す．その後，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することで新たな画像を生成し，NeRF の再学習に利用する手法について述べる．

3.1 Neural radiance field (NeRF)

NeRF [13] とは，3 次元空間中のある 1 点と方向を入力して，その点における密度と入力した方向における輝度を出力するニューラルネットワークである．一般的には視点位置が既知の多視点で撮影された画像を学習データとして，ボリュームレンダリングによってそれらを再構成するようにパラメータを学習する．学習後は任意の視点位置における画像を生成可能である．また，ボリュームレンダリングに使用した重みを用いて，擬似的に深度画像を生成することも可能である．

3.2 学習ベース両眼ステレオの NeRF 生成画像への適用

本研究は，学習ベース両眼ステレオがもつ事前知識を NeRF の学習に利用することを試みる．そこで，学習後の NeRF が生成したステレオ画像に対して，学習ベース両眼ステレオを適用する．3.1 で述べたように，学習後の NeRF を用いることで任意視点の画像を生成することができる．本研究では両眼ステレオに入力するステレオペアとして，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する．このステレオペアを学習済みの両眼ステレオに入力して視差を推定する．

図 1 に学習ベース両眼ステレオの一つである RAFT-Stereo [10] を適用した例を示す．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．また，NeRF がレンダリングした深度も合わせて示す．NeRF の既存手法である (a) K-planes [5] と (b) DäRF [19] について，DDP-NeRF [17] で用いられた ScanNet [2] 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という問題設定である．K-planes と DäRF の学習後にそれぞれステレオ画像を生成し，学習済み RAFT-Stereo を適用して学習に用いた視点における深度を推定した．図に示すように，NeRF がレンダリングした深度に大きく誤り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場合がある．

表 1 に定量評価を示す．各 1 行目は NeRF によってレンダリングした深度，2 行目は RAFT-Stereo により推定した深度の結果であり，評価指標には Eigen ら [4] を参考に，absolute relative error (AbsRel)， squared relative error (SqRel)，root mean squared error (RMSE)，root mean squared log error (RMSE log) を用いた．学習に用いた画像の枚数が少ないため，K-planes がレンダリングした深度には大きな誤差がみられる．これに対し，RAFT-Stereo により推定した深度は大幅に誤差が減少しており，学習ベース両眼ステレオは NeRF の新規視点合成におけるノイズに頑健であることがわかる．DäRF は学習に単眼深度推定の結果を利用する手法であり，学習時の幾何的拘束によってレンダリングした深度の誤差は小さいが，この場合においても学習ベース両眼ステレオによってさらに誤差が減少していることがわかる．

表 1 K-planes [5] と DäRF [19] について，ScanNet [2] 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 [10] を適用して推定した深度の定量評価

	AbsRel ↓	SqRel ↓	RMSE ↓	RMSE log ↓
K-planes [5]	0.410	1.172	1.449	0.520
K-planes [5] + RAFT-Stereo [10]	0.210	0.375	0.714	0.291
DäRF [19]	0.082	0.029	0.253	0.105
DäRF [19] + RAFT-Stereo [10]	**0.071**	**0.028**	**0.235**	**0.094**

以上により，NeRF が生成したステレオ画像に対して学習ベース両眼ステレオを適用することで，NeRF がレンダリングした深度に誤りが含まれている場合でも，視差を頑健に推定できることを示した．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．

3.3 手法の概要

本研究では学習ベース両眼ステレオで推定した視差を NeRF の学習に利用する手法を提案する．単純なアプローチとして，視差から計算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，推定した深度画像を直接用いるのではなく，学習に使用した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．図 2 に新たな学習画像を生成するプロセスを図示する．

(1) 既存の NeRF のモデルの学習を行う．
(2) 学習に用いた各視点で画像を再構成する．それに加えて，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) 学習に使用した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) 学習した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．

3.4 視差を用いた新たな学習画像の生成

本研究では最初に既存の NeRF の学習をした後，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．各視点における画像のペア (I_c, I_r) を学習ベース両眼ステレオ

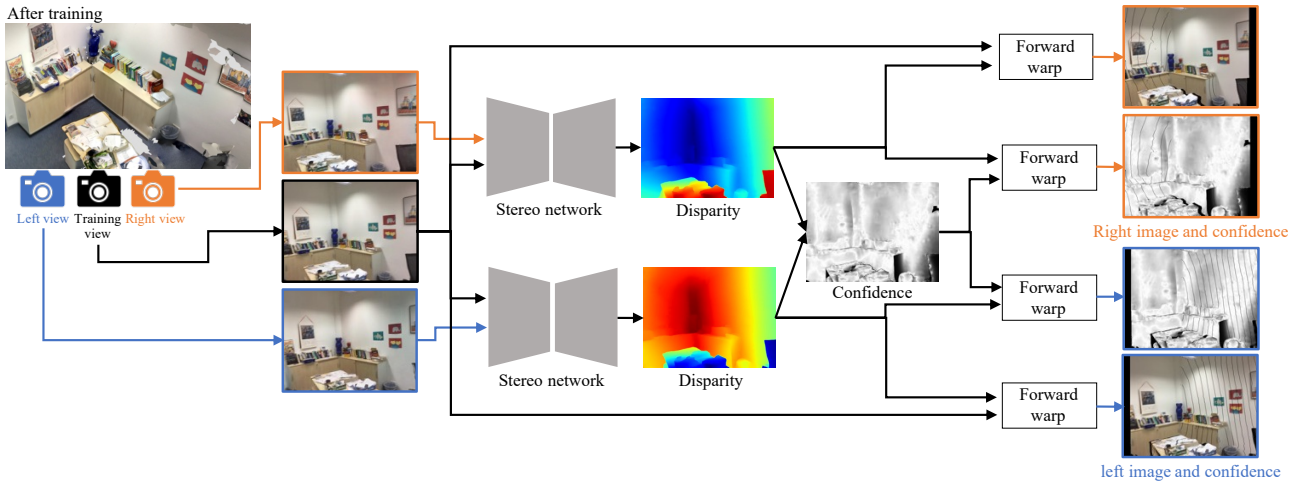
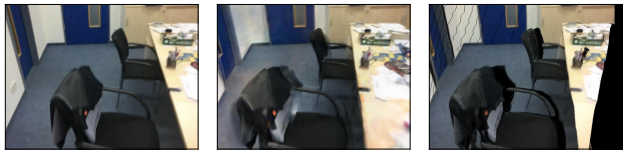


図 2 学習ベース両眼ステレオを用いた新たな学習画像の生成．NeRF の学習後，各学習視点に対して，視点位置を同じ基線長で左右にずらした画像を生成する．その後，学習視点で再構成した画像と左右で生成した画像をそれぞれペアとして，学習ベース両眼ステレオに入力し視差を推定する．推定した二つの視差から確信度を計算し，学習視点の画像と共に推定した視差で変形することで，新たな学習画像とその確信度を得る．



(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) 右方向にずらした視点で生成した画像，(c) 視差を用いて学習視点の画像を (b) の視点での画像に変形したもの．

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 は学習に使用した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，学習後はノイズの少ない画像が再構成される．(b) I_r は視点を右に微小にずらした視点で生成した画像である．学習データが少ない場合は，この例のように学習した視点から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．理想的には．NeRF の学習後，物体が存在しない空間では密度と輝度が 0 となるが，学習データが少ない場合にはそのような空間でも密度と輝度が値を持ってしまい，結果として雲のようなノイズや色の劣化が生じてしまう [3], [22]．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．

3.5 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，視点を右に微小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．したがって，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

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

で計算する．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．

3.6 誤差関数

生成した画像と確信度を用いて，以下の誤差関数を用いて NeRF の再学習を行う．

$$\mathcal{L} = \mathcal{L}_{\text{nerf}} + \mathcal{L}_s \qquad (3)$$

表 2 ScanNet [2] と Tanks and Temples [9] 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 K-planes [5] と DäRF [19] に適用したものがそれぞれ K-planes + stereo, DäRF + stereo である．各指標で最も良いものを**太字**，二番目に良いものを下線で示してある．DDP-NeRF [17] は SCADE [22] の論文中の out-of-domain の結果である．

	ScanNet [2]			Tanks ans Temples [9]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Vanilla NeRF [13]	19.03	0.670	0.398	17.19	0.559	0.457
NerfingMVS [28]	16.29	0.626	0.502	-	-	-
RegNeRF [14]	18.93	0.676	0.450	-	-	-
DS-NeRF [3]	20.85	0.713	0.344	-	-	-
DDP-NeRF [17]	19.29	0.695	0.368	19.18	0.651	0.361
SCADE [22]	21.54	0.732	**0.292**	20.13	0.662	0.358
K-planes [5]	18.80	0.715	0.385	17.27	0.600	0.423
K-planes [5] + stereo (Ours)	19.81	0.738	0.346	19.20	0.656	0.359
DäRF [19]	21.37	0.764	0.321	19.87	0.673	0.367
DäRF [19] + stereo (Ours)	**22.08**	**0.777**	0.305	**20.23**	**0.690**	**0.350**

表 3 誤差関数中の右の画像（$\hat{I}_r$），左の画像（$\hat{I}_l$），確信度（$\hat{C}$）に対する ablation study.

$\hat{I}_r$	$\hat{I}_l$	$\hat{C}$	PSNR ↑	SSIM ↑	LPIPS ↓
✓			21.92	0.774	0.312
✓	✓		21.90	0.776	0.309
✓		✓	22.00	0.775	0.310
✓	✓	✓	**22.08**	**0.777**	**0.305**

表 4 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
	PSNR ↑	SSIM ↑	LPIPS ↓	RMSE log ↓
$\lambda_d = 1.0$	21.42	0.762	0.321	0.111
$\lambda_d = 0.1$	21.73	0.774	0.307	0.103
$\lambda_d = 0$	**22.08**	**0.777**	**0.305**	**0.102**

ここで，$\mathcal{L}_{\text{nerf}}$ は元のモデルで用いられている誤差関数である．$\mathcal{L}_s$ は新たに生成した画像を用いた誤差関数であり以下で定義する．

$$\mathcal{L}_s = \frac{1}{2}\sum_{x,y}\Big(\hat{C}_r(x,y)(\hat{I}_r(x,y) - I_r^*(x,y))^2 + \hat{C}_l(x,y)(\hat{I}_l(x,y) - I_l^*(x,y))^2\Big) \quad (4)$$

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

4. 実験

4.1 実装

本研究では K-planes [5] と DäRF [19] について提案手法を適用した．また，学習ベース両眼ステレオについては RAFT-Stereo [10] の学習済みモデル（Sceneflow データセット [12] での事前学習と Middlebury データセット [18] でのファインチューニング）を用いた．3.3 で述べたように，それぞれのモデルを学習したのちステレオペアを生成し，RAFT-Stereo で視差を推定する．推定した視差を用いて画像を変形し，3.6 で述べた誤差関数によって各モデルの再学習を行う．

4.2 データセット

室内のシーンで撮影された二つのデータセット（ScanNet [2]，Tanks ans Temples [9]）を用いて実験を行なった．それぞれ DDP-NeRF [17]，SCADE [22] 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．

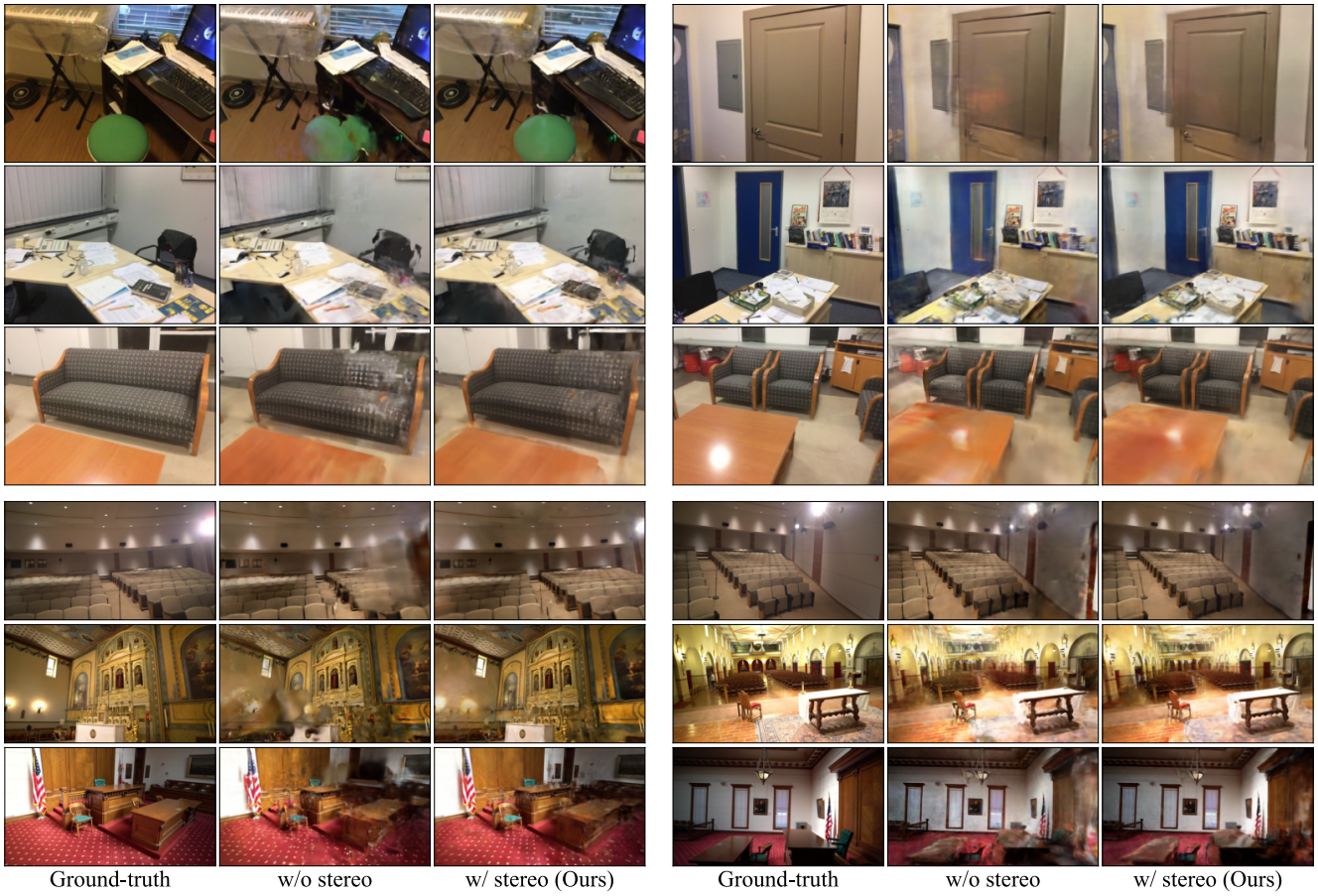
4.3 実験結果

新規視点合成

表 2 に新規視点合成の実験結果を示す．指標は PSNR, SSIM [27]，LPIPS [34] である．K-planes [5] と DäRF [19] の結果は我々で再学習を行なったものである．それ以外の比較手法の Vanilla NeRF [13], NerfingMVS [28], RegNeRF [14]，DS-NeRF [3]，DDP-NeRF [17]，SCADE [22] は，DDP-NeRF，SCADE，DäRF の論文で報告された値を示した．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．

表に示すように，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．3.4 で述べたように，提案手法



(a) K-planes [5]　　(b) DäRF [19]

図 4 ScanNet [2] と Tanks and Temples [9] データセットにおける新規視点合成の定性評価. (a) K-planes [5] と (b) DäRF [19] に対して，提案手法を適用したものは w/o stereo (Ours) で示してある.

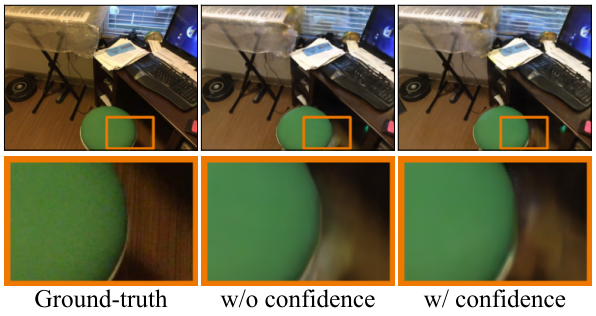


図 5 確信度有り無しでの比較．左右の一貫性を確信度として用いることにより，隠蔽により両眼ステレオの精度が低下する深度が不連続であるような箇所において誤差を低減できる．

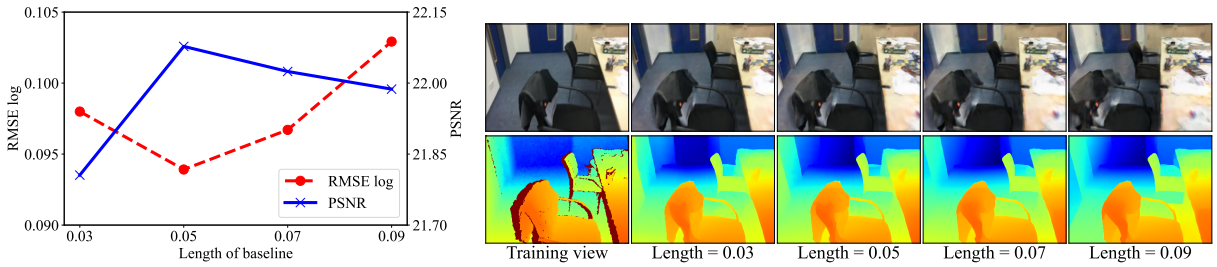
によって雲のようなノイズや色の劣化が低減できることが確認できる．

Ablation study

提案手法は式 (4) 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．これらに対する ablation study を表 3 に示す．各誤差関数の詳細は付録に記載している．この実験では DäRF [19] に対して提案手法を適用した．この表より，右方向のみを用いた場合や確信度を用いない場合に対する，左右方向の画像と確信度を用いた場合の有効性が確認できる．図 5 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．

基線長の影響

ステレオ画像を生成する際の基線長の影響について図 6 に示す．(a) は DäRF [19] ベースの提案手法に対して，基線長を $[0.03, 0.05, 0.07, 0.09]$ で実験した際の，RAFT-Stereo [10] で推定した深度の誤差（RMSE log）と，それらを用いて学習したモデルの新規視点合成の結果（PSNR）である．(b) には各基線長で生成したステレオ画像と，それらを入力して推定した深度画像を示してある．基本的にはどの基線長においても新規視点合成の精度が向上している．一方で，推定した深度画像の誤差と新規視点合成の結果の相関関係も見られる．一般的には大きな基線長を用いた方が，推定可能な深度の範囲が大きくなるが，(b) に示すように基線長を大きくすると NeRF が生成する画像のノイズや遮蔽による深度推定への悪影響も生じる．なお，本



(a) (b)

図 6 (a) 基線長を変えてステレオ画像を生成した場合の RAFT-Stereo [10] による深度推定の誤差（RMSE log）とそれらを用いた場合の新規視点合成の結果（PSNR）．(b) 左から学習に使用した視点で再構成した画像と正解の深度画像，各基線長で生成したステレオ画像とそれらから RAFT-Stereo が推定した深度画像．

研究ではすべての実験において基線長を 0.05 とした．

深度の誤差関数を加えた学習

両眼ステレオで得られた視差から深度が計算できるため，この深度に対する誤差関数を学習に加えることができる．表 4 に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．誤差関数の詳細は付録に記載している．ここで，λ_d は深度を用いた誤差に対する重みであり，$\lambda_d = 0$ は深度を用いない場合に対応する．この実験では提案手法を適用する手法に DäRF [19] を用いた．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，$\lambda_d = 1.0$ においては，レンダリングした深度の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．

5. 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．推定した視差を用いて学習視点で再構成した画像を変形し，新たな学習画像として再学習を行う手法を提案した．既存手法である K-planes [5] と DäRF [19] 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した．本手法は一般的な NeRF モデルすべてに対して適用可能である．課題としては，ステレオペア生成のために最初に NeRF の学習を行う必要があることが挙げられる．今後は，学習途中からステレオペアを生成したり，誤差関数を加えたりするなど，学習の高速化，最適化に取り組む予定である．

参考文献

[1] Chang, J.-R. and Chen, Y.-S.: Pyramid Stereo Matching Network, *CVPR* (2018).
[2] Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T. and Niessner, M.: ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes, *CVPR* (2017).
[3] Deng, K., Liu, A., Zhu, J.-Y. and Ramanan, D.: Depth-Supervised NeRF: Fewer Views and Faster Training for Free, *CVPR*, pp. 12882–12891 (2022).
[4] Eigen, D., Puhrsch, C. and Fergus, R.: Depth map prediction from a single image using a multi-scale deep network, *NeurIPS*, Vol. 2, pp. 2366–2374 (2014).
[5] 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).
[6] Jain, A., Tancik, M. and Abbeel, P.: Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis, *ICCV*, pp. 5885–5894 (2021).
[7] Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A. and Bry, A.: End-To-End Learning of Geometry and Context for Deep Stereo Regression, *ICCV* (2017).
[8] Kim, M., Seo, S. and Han, B.: InfoNeRF: Ray Entropy Minimization for Few-Shot Neural Volume Rendering, *CVPR*, pp. 12912–12921 (2022).
[9] 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).
[10] Lipson, L., Teed, Z. and Deng, J.: RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching, *3DV* (2021).
[11] Long, X., Lin, C., Wang, P., Komura, T. and Wang, W.: SparseNeuS: Fast Generalizable Neural Surface Reconstruction from Sparse Views, *ECCV* (2022).
[12] Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A. and Brox, T.: A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation, *CVPR* (2016).
[13] 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).
[14] Niemeyer, M., Barron, J. T., Mildenhall, B., Sajjadi, M. S. M., Geiger, A. and Radwan, N.: RegNeRF: Regularizing Neural Radiance Fields for View Synthesis From Sparse Inputs, *CVPR*, pp. 5480–5490 (2022).

[15] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. and Koltun, V.: Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer, *IEEE TPAMI*, Vol. 44, No. 03, pp. 1623–1637 (2022).
[16] Ranftl, R., Bochkovskiy, A. and Koltun, V.: Vision Transformers for Dense Prediction, *ICCV* (2021).
[17] 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).
[18] Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X. and Westling, P.: High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth, *Pattern Recognition* (Jiang, X., Hornegger, J. and Koch, R., eds.), pp. 31–42 (2014).
[19] 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).
[20] Tosi, F., Tonioni, A., De Gregorio, D. and Poggi, M.: NeRF-Supervised Deep Stereo, *CVPR*, pp. 855–866 (2023).
[21] Truong, P., Rakotosaona, M.-J., Manhardt, F. and Tombari, F.: SPARF: Neural Radiance Fields From Sparse and Noisy Poses, *CVPR*, pp. 4190–4200 (2023).
[22] Uy, M. A., Martin-Brualla, R., Guibas, L. and Li, K.: SCADE: NeRFs from Space Carving with Ambiguity-Aware Depth Estimates, *CVPR* (2023).
[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention is all you need, *NeurIPS*, p. 6000–6010 (2017).
[24] Wang, C., Sun, J., Liu, L., Wu, C., Shen, Z., Wu, D., Dai, Y. and Zhang, L.: Digging into Depth Priors for Outdoor Neural Radiance Fields, *ACM MM*, p. 1221–1230 (2023).
[25] Wang, G., Chen, Z., Loy, C. C. and Liu, Z.: SparseNeRF: Distilling Depth Ranking for Few-shot Novel View Synthesis, *ICCV*, pp. 9065–9076 (2023).
[26] Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T. and Wang, W.: NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction, *NeurIPS* (2021).
[27] Wang, Z., Bovik, A., Sheikh, H. and Simoncelli, E.: Image quality assessment: from error visibility to structural similarity, *IEEE TIP*, Vol. 13, No. 4, pp. 600–612 (2004).
[28] Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J. and Zhou, J.: NerfingMVS: Guided Optimization of Neural Radiance Fields for Indoor Multi-View Stereo, *ICCV*, pp. 5610–5619 (2021).
[29] Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Yu, F., Tao, D. and Geiger, A.: Unifying Flow, Stereo and Depth Estimation, *IEEE TPAMI*, Vol. 45, No. 11, pp. 13941–13958 (2023).
[30] Yariv, L., Gu, J., Kasten, Y. and Lipman, Y.: Volume rendering of neural implicit surfaces, *NeurIPS* (2021).
[31] Yu, A., Ye, V., Tancik, M. and Kanazawa, A.: pixelNeRF: Neural Radiance Fields From One or Few Images, *CVPR*, pp. 4578–4587 (2021).
[32] Yu, Z., Peng, S., Niemeyer, M., Sattler, T. and Geiger, A.: MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction, *NeurIPS* (2022).
[33] Zhang, F., Prisacariu, V., Yang, R. and Torr, P. H.: GA-Net: Guided Aggregation Net for End-To-End Stereo Matching, *CVPR* (2019).
[34] Zhang, R., Isola, P., Efros, A. A., Shechtman, E. and Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, *CVPR* (2018).

付　　録

A.1 誤差関数の詳細

A.1.1 Ablation study

表 3 の実験で用いた誤差関数は以下である．

$$\mathcal{L}_s = \sum_{x,y} (\hat{I}_r(x,y) - I_r^*(x,y))^2 \tag{A.1}$$

$$\mathcal{L}_s = \frac{1}{2} \sum_{x,y} \Big((\hat{I}_r(x,y) - I_r^*(x,y))^2 + (\hat{I}_l(x,y) - I_l^*(x,y))^2 \Big) \tag{A.2}$$

$$\mathcal{L}_s = \sum_{x,y} \hat{C}_r(x,y) (\hat{I}_r(x,y) - I_r^*(x,y))^2 \tag{A.3}$$

A.1.2 深度を用いた誤差

表 4 の実験で用いた誤差関数は以下である．

$$\mathcal{L} = \mathcal{L}_{\text{nerf}} + \mathcal{L}_s + \lambda_d \mathcal{L}_d \tag{A.4}$$

ここで $\mathcal{L}_d$ が両眼ステレオで得られた深度を用いた誤差であり，実験では確信度 C_c を用いて以下のように定義した．

$$\mathcal{L}_d = \sum_{x,y} C_c(x,y) |z_r(x,y) - z^*(x,y)| \tag{A.5}$$

z_r は視差 d_r から計算した深度（$z_r = bf/z_r$ で b，f はそれぞれ既知の基線長と焦点距離）であり，z^* は NeRF がレンダリングした深度である．