H-021

Finetuning-Less 視差推定に向けた NeRF を用いた学習データ生成の検討
Towards Finetuning-Less Disparity Estimation: An Investigation of Training Data Generation Using NeRF

中野　和香子† 谷村　崇仁† 城戸　英彰† 入江　耕太‡
Wakako Nakano Takahito Tanimura Hideaki Kido Kota Irie

1. はじめに

自動運転システムにおけるステレオ画像からの視差推定は，障害物検知や運転計画などの主要タスクを支える重要な要素技術である．近年，従来の幾何学的なステレオマッチング手法ではなく，深層学習を用いた視差推定手法により，各種ベンチマーク [1,2] で高い視差推定精度の達成が報告されている．

自動運転用途での視差推定モデルの学習には多様な走行シーンや車種ごとのカメラ取付け位置・視野角の違いに対応するため，多様かつ大量の学習データが不可欠であり，センサ・実車走行を含めたデータ収集コストが高い．

そのため，従来はゲームエンジンを利用したセンサシミュレーションにより得られる合成データでの学習が研究されている．[3,4] しかし，合成データではドメインシフトと呼ばれる生成画像の輝度分布や生成視差ラベルの距離分布と実データ分布との間に乖離が生まれ，これが視差推定精度劣化につながる．このため，合成データでの事前学習の後，複数のラベル付き実車データによるファインチューニングが必要である．

そこで近年，NeRF (Neural Radiance Fields) [5] を始めとするニューラルレンダリング技術が注目を集めている．NeRF は，少数の入力画像からシーンを高精度に再構成し新たな視点画像を生成可能である．この特性を用い，NeRF から視差推定の学習データ生成する研究が行われている．上記研究では，様々なドメインのアプリケーションで視差推定精度が向上することが示されている．[11] しかし，カメラポーズ推定誤差は生成視差ラベルへの影響が大きく，また路面のようなローテクスチャな領域の精度が低いため，自動運転のアプリケーションへの適用に向けては精度が不十分であり，より高精度な車両シーンのデータが必要である．

また，より高精度な車両シーン向けの NeRF 手法として，車両のマルチカメラ画像と LiDAR を組み合わせた屋外シーンを学習する手法 [7,8] が提案されているが，これらの視差推定タスクへの適用検討および評価は十分に行われていない．

そこで本研究では，図 1 に示すような単一車種でのセンサデータのみを用い，車種やカメラ取付け位置に対応した学習データ生成を可能とすることでファインチューニングを不要とする”Finetuning-Less”な視差推定をめざし，車両向けの NeRF 手法を用いた視差推定フレームワークを提案する．単一車種取得データから他車種への適用時の評価を実施するため，評価実験では Waymo Open Dataset [9] を用いて学習した NeRF から KITTI15 互換のデータを生成，学習し視差推定精度を測定することで、自動運転向けのアプリケーションのデータセットにおける有効性を評価する．

† 日立製作所 研究開発グループ　Hitachi, Ltd. R&D Group
‡ Astemo　Astemo, Ltd.

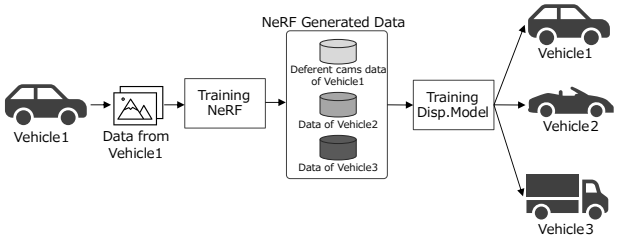

図 1　Finetuning-Less 視差推定に向けたデータ生成

2. 関連研究

2.1 視差推定ネットワーク

近年，ステレオ視差推定は，エンドツーエンドの深層学習により大幅な性能向上が報告されている．例えば，RAFT-Stereo [13] や IGEV-Stereo [14] のような大型モデルは自動運転向けのベンチマーク KITTI15 において最先端な推定精度を達成している．一方，組み込み GPU や車載 SoC 上での高速な動作のために，パラメータ量の少ない軽量モデルも提案されている．[10]

これらの手法は，いずれも高い精度達成のために大規模かつ正確な正解ラベルの収集を必要とする.しかし実環境で LiDAR とカメラを複数車両に搭載し，キャリブレーションを行いながらラベルを収集するには多大なコストがかかる．そのため，データ生成により，データ収集コストを削減する研究が行われている．

2.2 生成データによる学習

2.2.1 ゲームエンジンを用いた学習データ生成

合成データを大量に生成し学習に用いる手法として，ゲームエンジンを用いた車両センサの出力と制御情報のシミュレーションが研究されている [1,3]．これらは，多様な環境・センサ条件を効率的に生成可能である．一方で，合成データと実車データ間でのデータ分布差によるドメインシフトが問題となり，合成データだけで学習したモデルを実環境に適用すると性能が低下する [12]．従って，合成データでの事前学習の後，少量の実データでファインチューニングを行うアプローチが一般的である．

2.2.2 NeRF による学習データ生成

ゲームエンジンに代わる手段として NeRF（Neural Radiance Fields） [5] によるニューラルレンダリングが注目を集めている．NeRF は 3 次元空間内のある座標と視線方向を入力とし，その座標における輝度と密度を推定するニューラルネットワークである．少数の実画像から写実的な新

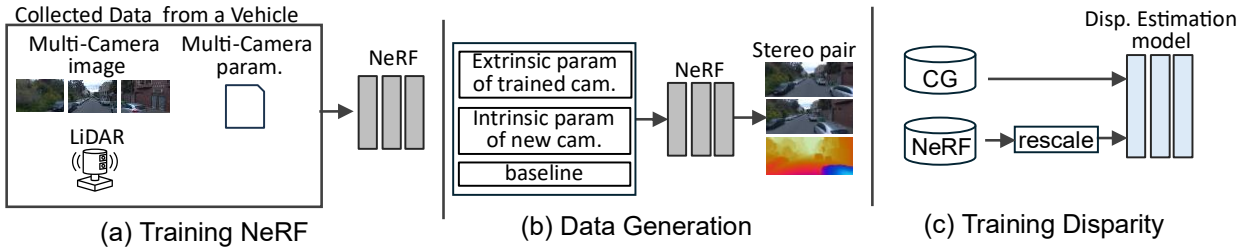

図 2　NeRF による視差推定学習データの生成

規視点画像と対応する視差ラベルを生成できるため，視差推定モデルへの応用が検討されている．[11] しかしこの手法は単眼画像のみで NeRF を学習する方式を用いており，カメラ姿勢推定の誤差が，直接生成視差に影響し精度の低下を招く．また，屋内シーンや 360° シーンを対象とした NeRF の使用により，テクスチャの乏しい領域で精度が低下することが指摘されている．[6] これらの要因により，自動運転のアプリケーションへの適用に向けてはゲームエンジンベースのデータ生成手法とほとんど同等の性能であり精度が不十分である．精度向上には，より高精度な車両シーンのデータ生成が必要である．

最近の研究では，高精度な車両シーンの NeRF 手法として，車載マルチカメラと LiDAR 情報を統合し NeRF を学習することで，屋外シーンでの再構築精度を向上させる手法が提案されている [7,8]．これらの手法は，LiDAR により幾何拘束を補完することで，再構成精度を向上させているが，自動運転アプリケーション向けの視差推定への適用においては，検討・評価が十分行われておらず，データ多様性のおよび実環境との輝度分布の差が問題となる可能性がある．

ゲームエンジンによるデータ生成はドメインギャップ，既存の NeRF によるデータ生成は車両シーン生成と深度精度の不足という課題を残している．そこで本研究では，そこで本研究は車載マルチカメラ画像と LiDAR 点群を利用した NeRF により学習データを生成、視差推定モデルを学習するフレームワークを提案する．従来のゲームエンジン生成データ（SceneFlow）を用いた学習との比較評価を行う．従来手法との位置づけは表 1 に示す．

3. 提案手法

本研究では，“Finetuning-Less” 視差推定の実現に向け，車載マルチカメラ画像と LiDAR 点群を利用した NeRF により視差推定学習データを生成するフレームワークを提案する．3.1 節では，フレームワークの概要を 3.2 節では，NeRF からの視差学習データ生成の詳細を説明する．

3.1 データ生成フレームワークの概要

本手法で提案する視差推定学習データを生成するフレームワークの概要を図 2 に示す．フレームワークは，高精度な車両シーンを再構成するための NeRF 学習 (a)，NeRF から視差推定学習データを生成するデータ生成 (b)，生成データを用いて視差推定モデルの学習を実施する視差学習 (c)のフェーズからなる。

NeRF 学習：単一車両に搭載したマルチカメラの画像，カメラの内部・外部パラメータ，LiDAR 点群を走行により収集。前記収集データを用いて NeRF を学習する．本研究では，NeRF 手法として，カメラ画像と LiDAR 情報の両方を利用して車両周辺環境を高精度に再構築可能な StreetSurf を採用した．[7]

データ生成：高精度かつ必要となる任意の車両カメラに適合したデータ生成をするため，学習済み NeRF に対し，任意のステレオカメラ内部パラメータと，基線長，NeRF 学習時の外部パラメータを設定してステレオ画像および深度情報を生成する．前記深度情報から視差ラベルを算出することで，単一車両の収集データから，他車種等の多様なカメラ配置を想定した視差学習データセットの生成が可能となる．

視差学習：生成したステレオ画像および視差ラベルを従来のゲームエンジンによる生成データに加えることで併用して視差推定モデルの学習を行う．この時，NeRF による生成データから実環境の特徴を反映しつつも，NeRF 生成時のノイズの影響を低減するため、スケーリングをずらし学習を実施する．

3.2 視差学習データの生成

データ生成フェーズにおける，ステレオ画像および視差ラベル生成の詳細を説明する．

生成データは，必要となる任意の車両カメラに適合させるため，NeRF から生成する画像のカメラ内部パラメータは新しく必要とするカメラの内部パラメータを設定する．

外部パラメータは，NeRF 学習時のカメラ姿勢や撮影領域に近い範囲ほど再構成精度が高くなることが知られているため，ステレオカメラのうち視差を推定するカメラに対応する撮影座標を，NeRF 学習時に使用したカメラの座標と設定する．また，もう一方のカメラはそこから基線長bだけ水平方向に移動させた座標に設定し，画像および深度を生成する．

また，前記の NeRF から出力される深度ラベルを用いて視差ラベル d を計算する．深度値を Z，基線長を b，焦点距離を f とすると，視差 d は以下の式で表される．

表 1 従来手法に対する本研究の位置づけ

Method	Vehicle data	Camera flexibility	Extra sensor	Main issue	Finetuning requirement
SceneFlow (synth) [3]	✓	✓	None	Domain gap to real images	High
NeRF-Sup. Stereo [11]	✗	✓	None	No vehicle scenes; depth noise	High
This work	✓	✓	One-off LiDAR	One-off LiDAR acquisition	Reduced

$$d = \frac{(b \cdot f)}{\mathrm{Z}}$$

NeRF が出力する深度値 Z は通常，モデル内部で任意のスケールをもつため，本研究では Z を 0 から 1 の範囲に正規化し，最大距離で再スケールしてから上式を適用する．また，天空領域など極端に遠方にある要素の深度値では視差値が 0 になるよう処理する．

4. 実験・評価

本提案のフレームワークの有効性を，本フレームワークと，従来手法であるゲームエンジンを用いた生成データを用いた学習と比較して検証する．

特に，従来手法による車両シーンの生成データと，本フレームワークでの NeRF による車両シーンの生成データでの学習効果を比較するため，ゲームエンジンによる生成データとして SceneFlow を用い，そのうち車両シーンの生成データを本フレームワークによる生成データで置き換えたカスタムデータセットを作成し比較した．

このとき，単一車両取得データから他車両ステレオカメラへの適用時の評価を実施するため，評価実験では Waymo Open Dataset [9]を用いて StreetSurf を学習し，KITTI15 互換のデータを生成した．

また，ゲームエンジンによる生成データと本フレームワークでの生成データの混合比率の観点から精度に及ぼす影響を評価した．

4.1 評価設定

すべての実験は NVIDIA A100 GPU 1 台上で実施し，本実験の主要な設定として生成データの条件，視差推定モデルの学習条件，そして比較対象とした ゲームエンジン合成データの設定について述べる．

NeRF 生成データ条件：StreetSurf を用い，Waymo Open Dataset [9] 内の 24 シーンから RGB, LiDAR, 法線を学習．各シーンでは，学習時とは異なる KITTI15 でのカメラ内部パラメータ（焦点距離，カメラ歪み）および 基線長を適用してカメラステレオペアおよび深度情報を生成した．視差算出は，深度最大値は 120 として算出した．また，各シーンでは，初めの 5 ペアおよび末尾 10 ペアは除外し，合計 7560 ペアの合成データを作成した．

視差推定モデルの学習・評価条件：視差推定モデルには，車載向けの軽量モデルである FADNet++ [10]を用い，64 エポック，バッチサイズ 4 で学習を実施した．評価は，KITTI15 データセットの Train データ 200 枚から，ランダムに選択した 180 枚をテストデータに用いた．

ゲームエンジン生成データの設定：比較対象として使用したゲームエンジン生成データには SceneFlow [3]を採用している．SceneFlow は多様な三次元特徴を学習するために作成されたデータセットであり，ランダムな 3D 軌道に沿って飛行する日常的なオブジェクトで構成される FlyingThings3D 22390 データ，アニメーションシーンで構成される Monkaa 8664 データ，KITTI15 に近いデータを含む車両シーンで構成される Driving 4400 データで構成されている．

表 2　Sceneflow および提案の視差誤差比較

	D1-all(%) ↓	EPE ↓
SceneFlow	11.17	2.75
This work (mixed-scale)	8.54	2.07
This work (uniform-scale)	83.1	11.8

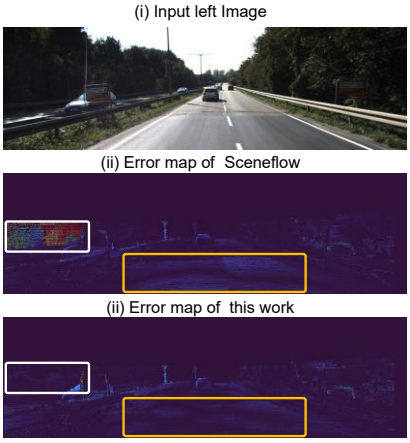

図 3　生成データにおける視差誤差比較

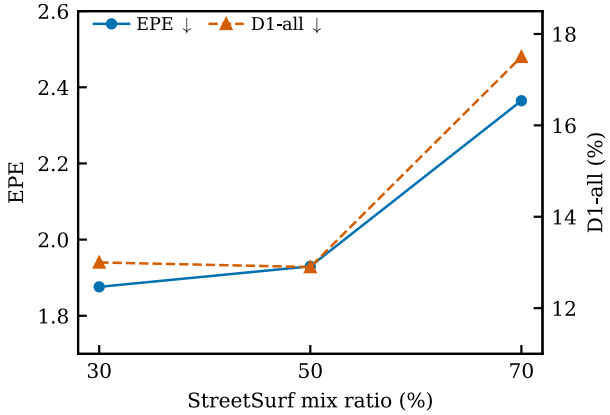

図 4　生成データ比率による精度比較

4.2　生成データの精度比較評価

本評価では，提案手法での生成データと従来のゲームエンジンを用いた車両シーンの生成データの比較をするため StreetSurf を用いたカスタムデータセットを作成し，SceneFlow での学習時との比較評価を実施した．

カスタムデータセットは，SceneFlow を構成する FlyingThings3D，Monkaa，Driving データのうち 14%に当たる車両シーンの Driving データを，NeRF 生成データからランダムサンプリングした同量のデータで置換し作成した．

また，カスタムデータセットは，学習精度の向上を目的に輝度値を意図的にずらしたケースについても評価を行った(mixed-scale)．ここでは，下記の 2 パターンでステレオ画像の輝度値をスケーリングし評価した．

① mixed-scale:NeRF 生成データの輝度レンジを混合する SceneFlow の FlyingThings3D, Monkaa データセットより 1/255 に設定し，別のスケールで学習

② uniform-scale: StreetSurf 生成データは輝度レンジをこれと混合する FlyingThings3D, Monkaa と同じ輝度レンジで学習

定量評価として，SceneFlow データセット, カスタムデータセットそれぞれで学習した結果を表 2 に示す．D1-all は 5%および 3 ピクセル以上の誤差が生じたピクセルの割合，EPE は平均の視差誤差を示している．カスタムデータセットにおいては，mixed-scale の場合に SceneFlow の学習精度を上回り， D1-all の精度が 23%改善した．

また，カスタムデータ学習時の輝度スケールの比較では，輝度スケールを意図的にずらしたデータ混合が，FADNet++ における視差推定性能を飛躍的に向上させることが明らかになった．

定性的な視覚評価として，推論時に出力された視差ラベルにおける誤差を図 3 に示す．上から，入力左画像，SceneFlow での学習モデルにおける推論誤差マップ，カスタムデータセットでの学習モデルにおける推論誤差マップを示している．推論誤差は正解視差と推論視差の差分であり，紺から赤色となるほど視差誤差が大きい (水色＜緑＜黄＜橙＜赤の順)．カスタムデータセットでは，特に橙枠で囲んだ道路部分や白枠で囲んだ樹木，といった車両走行環境特有の背景に対する認識精度の向上が確認された．一方で，物体の背景との境界付近においては，SceneFlow と同等の誤差が生じている．

4.3 データ比率の影響評価

本提案のフレームワークによる生成データの影響をより詳細に調べるため，NeRF 生成データの総学習データに占める比率が視差推定精度に与える影響を，以下の条件で検証した．

NeRF 生成データの比率を 30%，50%，70%と変化させ，残りは SceneFlow から Driving データを除いた FlyingThings3D および Monkaa を用いて総サンプル数を一定に維持し学習した．

結果を図 4 に示す．D1-all は StreetSurf 比率が 50%の場合に最良の値をとる一方で，EPE は誤差が単調増加した．また 70%では両指標とも精度が大きく劣化した．本提案データを適度に混合することは外れ値が抑制されるものの，比率を過度に高めると生成視差ラベルの品質が，学習結果に影響を与え精度劣化させることが示唆される．

5. 結論

本研究では，ステレオ画像と視差ラベルを自動生成し，追加の Finetuning を行わずに視差推定モデルを学習する“Finetuning-Less”視差推定に向け，車載マルチカメラ画像と LiDAR 点群から学習した NeRF を用い，単一車種から多車種・多カメラ配置に対応する視差推定向け学習データ生成フレームワークを提案した。

本提案により，Waymo Open Dataset で学習した StreetSurf から KITTI15 向けのデータ生成を実施し，車載向けの小型視差推定モデルを対象とし従来のゲームエンジンでの車両シーンの生成データでの学習との比較評価を実施した結果，精度が 23%向上した．

以上の結果から，提案フレームワークが 1 台分の車両取集データだけでファインチューニングの必要性を低減することを確認した．

しかし，定性的評価では，車両走行環境特有の背景（路面，樹木）で大幅な誤差低減が見られた一方，物体境界付近の誤差は残存した。また，提案手法による生成データを過度にデータセットに混合することで精度低下を招くことがわかった．今後は真に “Finetuning-Less”な視差推定に向けては，生成視差ラベルに対する詳細な分析およびラベルの視差誤差低減の検討を実施する予定である．

参考文献

[1] M. Menze and A. Geiger, “Object scene flow for autonomous vehicles,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition. (CVPR), 2015.
[2] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, “A multi-view stereo benchmark with high-resolution images and multi-camera videos,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition., pp. 3260–3269, 2017.
[3] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, “Practical stereo matching via cascaded recurrent network with adaptive correlation,” Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition., pp. 16263–16272, 2022.
[4] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, “A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Jun. 2016.
[5] 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,” Proc. ECCV, 2020.
[6] M. K. Gjerde, F. Slezák, J. B. Haurum, and T. B. Moeslund, “From NeRF to 3DGS: A Leap in Stereo Dataset Quality?,” Proc. Synthetic Data for Computer Vision Workshop, CVPR, 2024.
[7] J. Guo, N. Deng, X. Li, Y. Bai, B. Shi, C. Wang, C. Ding, D. Wang, and Y. Li, “StreetSurf: Extending multi-view implicit surface reconstruction to street views,” arXiv preprint arXiv:2306.04988, 2023.
[8] Z. Wang, T. Shem, J. Gao, S. Huang, J. Munkberg, J. Hasselgren, Z. Gojcic, W. Chen, and S. Fidler, “Neural Fields meet Explicit Geometric Representations for Inverse Rendering of Urban Scenes,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2023.
[9] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al., “Scalability in perception for autonomous driving: Waymo open dataset,” Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition., pp. 2446–2454, 2020.
[10] Q. Wang, S. Shi, S. Zheng, K. Zhao, and X. Chu, “Fadnet++: Real-time and accurate disparity estimation with configurable networks,” arXiv preprint arXiv:2110.02582, 2021.
[11] F. Tosi, A. Tonioni, D. De Gregorio, and M. Poggi, “NeRF-Supervised Deep Stereo,” Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2023.
[12] F. Aleotti, F. Tosi, P. Z. Ramirez, M. Poggi, S. Salti, L. Di Stefano, and S. Mattoccia, “Neural disparity refinement for arbitrary resolution stereo,” Proc. Int. Conf. on 3D Vision (3DV), 2021.
[13] L. Lipson, Z. Teed, and J. Deng, “RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching,” Proc. Int. Conf. on 3D Vision (3DV), 2021.
[14] G. Xu, X. Wang, X. Ding, and X. Yang, “Iterative Geometry Encoding Volume for Stereo Matching,” Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2023.