Vol.2022-CVIM-228 No.20
2022/1/28

走査型 LiDAR のための時刻情報を用いた超解像

林佑介[1]　田口賢佑[1]　森田渉吾[1]　今枝航[1]　藤吉弘宣[2]

概要：LiDAR は広い範囲にわたって高精度に距離や位置，形状を検出し，3 次元での把握が可能である．その反面，カメラよりも解像度およびフレームレートが低いという課題がある．また，レーザー光をスキャンする走査型 LiDAR では，各 Depth 測定値の測定時刻が異なり，この測定時刻の違いに起因して移動物体の Depth がずれるという課題がある．本研究では，高解像度で移動体に対してずれのない Depth を得るため，走査型 LiDAR で得られた Depth 測定値を時間方向に拡張した 3 次元の Voxel 表現とし，これを入力データとしてニューラルネットワークを用いて推定する．空間的には遠方や物体の境界付近，時間的には目的の時刻に遠い時刻や移動物体のように高解像度化が困難な Depth 測定値に対して，背景と移動体の切り分けや不確実性を考慮することで精度向上を図る．

キーワード：走査式 LiDAR，超解像，不確実性，CG データ

1. はじめに

突風による道路への落下物や集中豪雨による土砂の流入の迅速な検知など，屋外への常時設置が必要となるインフラ監視において，より正確に状況を把握するために LiDAR は有用なセンシングデバイスの一つとなる． レーザー光を反射させ，戻ってくるまでの時間を計測する LiDAR は，広い範囲にわたって高精度に距離や位置，形状を検出し，3 次元での把握が可能になる．その反面，カメラよりも解像度およびフレームレートが低いという課題がある．また，レーザー光の照射方法は大きく分けて走査型とフラッシュ方式の 2 つがあるが，フラッシュ方式は面で光源を照射するため検出距離が走査型よりも短くなり，外光で測定結果がばらついてしまうために屋外での使用にも向いていない．一方，走査型 LiDAR は受光面積当たりのレーザーの光量を大きくできるために検出可能距離を長くすることができるのがメリットであるが，モーターを用いた回転機構や MEMS 等によりレーザー光をスキャンする必要があるため，各 Depth の測定時刻が異なる．この測定時刻の違いに起因して，車やバイクなどの移動物体の Depth が実際の値とずれてしまうという課題がある．

本研究は，走査型 LiDAR においてレーザー光のスキャンによる時刻ずれを解消し，移動体に対してずれのない正確な Depth を得るための超解像技術として以下の 3 つの項目を提案する．

①高解像度で移動体に対してずれのない Depth を得るため，走査型 LiDAR で得られた Depth 測定値を時間方向に 3 次元に拡張した Voxel 表現で表し，これを入力データとして全層畳み込みネットワークで推定

②畳み込みが有効に特徴を把握できるようにするため，画像から背景を抽出するマスクを作成することで時空間上の Depth 測定値を高密度化

③時空間方向の不確実性を考慮することで移動体の Depth 推定の精度向上

ただし，本研究で用いる LiDAR はカメラとの光軸が一致したフュージョンセンサで，Depth 測定値と画像の視差がない．また，本手法はこのフュージョンセンサを固定して設置した場合に限定し，静的なシーンの Depth マップの超解像を扱う．

2. 関連研究

2.1 超解像

超解像には 1 枚のフレームを用いるものと複数フレームを用いるものの 2 種類がある．1 枚のフレームの超解像としては，ニューラルネットワークを超解像に適用し end-to-end で学習したモデル[1]がある．また，[2]では ResNet[3]の構造を用いて入力層と差分値のみを学習し，スキップ接続により入力情報を保持して復元するという提案がされている．複数のフレームを用いた超解像としては，現フレームと前フレームを入力し，出力として各ピクセルがどの方向に動いたかを推定したフローマップを使って前フレームの特徴量マップを現フレーム用に変換する提案[4]がある．また，1 枚の画像を時関方向に拡張した Voxel 表現とし，コード化した露出と組み合わせることにより時空間上での超解像を行う研究[5]がある．Voxel 表現を用いた別の研究として，[6]ではイベントカメラを用いて疎なイベント情報を時空間上に表し，リカレントニューラルネットワークを用いて高速高解像なビデオを生成する．本研究では，これらの Voxel 表現を用いた研究と同様に，Depth 測定値を時空間上に表すとともに背景をマスクにより抽出することで高密度化することで超解像を行う．

2.2 不確実性

不確実性の推定は，ベイジアンニューラルネットワーク[7]をはじめとして，コンピュータビジョンを対象としたニ

1 京セラ（㈱）先進技術研究所
Advanced Technology Research Institute, Kyocera Corp.
2 中部大学
Chubu University

ューラルネットワークの研究においても様々な報告がなされている. ベイジアンニューラルネットワークにおいては，重みの分布から異なるネットワークを複数回サンプリングすることにより，ターゲットとなる分布の平均と分散を算出する．[8]では，サンプリングが変分推論に置き換えられている．別の方法としては，アンサンブル[9]やモンテカルロドロップアウト[10]が用いられる．異なるアプローチとしては, 推論の際に不確実性を算出する研究[11]がある. 不確実性を応用する研究として，焦点ずれによるレンズのぼけがニューラルネットワークに与える影響を解析するためにモデルの知識不足か否かを示す認識論的不確実性（Epistemic Uncertainty）を分析した研究[12]がある. 本研究では[11]の手法を用いて学習の際に Depth 推定するとともにデータの難しさを表す偶然的不確実性（Aleatoric Uncertainty）を算出することで Depth の超解像の性能が向上することを検証する.

2.3 CG データ

運転シーンをシミュレーションした CG データとして SYNTHIA[13]や VirtualKiTTI[14]がある．SYNTHIA では画像にセマンティックセグメンテーションのラベルが付けられている一方，VirtualKITTI では KITTI[15]の動画がシミュレーションされ，画像に完全に一致した Depth が画素ごとに対応付けされている．CARLA[16]は自動運転のアルゴリズム開発，検証を支援するためのシミュレーターである．カメラや LiDAR のセンサ出力に対応しており，シミュレーター空間上の任意の位置にセンサを設置することが可能である．しかしながら，これらのデータセットではフレームレートはいずれも遅く，1 フレーム分の時間内の各時刻に対応する教師データとなる Depth 情報を得ることができない. そこで本論文では CARLA でも使用されている 3D シミュレーターの Unreal Engine4[17]と 3D CG データベースとして AUTOCity[18]を用いて CG データを作成した.

2.4 ローリングシャッターおよび LiDAR ノイズの解消

従来の LiDAR に関する研究では，Depth 測定値はノイズがないものとされている．しかしながら，前述した KITTI のような公開データセットにおいても，たとえエゴモーションを見込んでローリングシャッターの影響を補正しても，透明物体や反射強度の高い物体においてノイズが生じることが報告されている[21]．3D モデルを用いて人間が目視でこれらのノイズを除去することは可能だが，アノテーションコストが非常に高い．カメラのローリングシャッターを解消する研究に注目すると，[19]では平面運動を仮定して画像の歪みを補正する方法が提案されている．また，列ごとの読み出しタイミングと露出の長さを制御して 3 次元の時空間のシーン情報をサンプリングすることで，高速撮影や動きブレのない画像生成の提案[20]がなされている．LiDAR の Depth 測定値のノイズ除去に関連する研究として，LiDAR のローリングシャッターの効果とキャリブレーションミスによる Depth 測定値のずれに対して，[21]ではステレオカメラの輝度情報の整合性（Photometric Consistency）およびステレオカメラと LiDAR の Depth 情報の整合性（Depth Consistency）を用いてノイズ除去を行う提案がされている．[22]では，Depth Completion のタスクに対して，実空間上のノイズを含んだ疎な Depth を CG 空間上で再現するドメイン適応の手法が提案されている．本研究では LiDAR の Depth 測定値に含まれるノイズがないよう，教師データを CG 上で作成することによってノイズ除去された高解像度な Depth 推定が可能であることを示す.

3. 提案手法

3.1 概要

我々の目的は，1 フレーム $F=\{f_t\}\ \ for\ t\in[0,T-1]$ の Depth 測定値からある時刻の Depth マップを推定することである．今回の検証では，推定する Depth マップの時刻は 1 フレームのうちの中央値の時刻${}^{T}/_{2}$とする. また，推定する Depth マップの解像度は元の 1 フレームで得られる解像度よりも高い解像度とする．この超解像のタスクを達成するために, 1 フレームの Depth 測定値を時空間の Voxel として表現し，この Voxel を入力として全層畳み込みネットワークで推定を行う．ネットワークの学習は教師あり学習の手法を使い，学習データとして大量の CG で作成した距離データを用いる.

3.2 入力データの表現方法及びマスクによる前処理

$F(x,y,t)$は$M\times N$ピクセルと 1 フレーム分の時間 T に対応する時空間の体積を表現した Voxel である．従来の LiDAR は, この Voxel を時間軸上に投影し, その結果$M\times N$の解像度を得る. 我々は, 時空間分解能でL倍のゲインを達成したい，つまり，$M\times N\times L$の時空間体積を実現する超解像を行いたい.

全層畳み込みネットワークが Voxel を効率よく処理できるようにするため，各時刻 t に対応した背景を抽出するマスク$M(x,y,t)$と 1 フレームの累積との要素積を算出し，Voxel の背景部分はこの要素積となるよう重ね合せる．これにより背景の Depth を高密度化する. 各時刻のマスクの作成は, 例えば LiDAR の 1 フレーム分の時間で得られる画像 2 枚を用いてオプティカルフローを算出し，等速直線運動を仮定して各時刻の背景を抽出する. 本研究の実験では, CG データで作成したマスクを用いた.

3.3 学習データ

我々のネットワークは 1 フレーム分の時間内の各時刻に対応する Depth 測定値が教師データとして必要である．しかしながら，そのような Depth 測定値および画像を含む公開データセットはない. そのため, CG で作成した画像および Depth を用いて学習することを提案する.

我々は，3D シミュレーターとして Unreal Engine4 を，3D CG データベースとして AUTOCity を用いた. AUTOCity は

物理ベースのレンダリングが可能で，カメラで撮影される画像と LiDAR で得られる Depth 情報を作成でき，仮想的に光軸を一致させることできる．そのため，我々が想定しているカメラと LiDAR の光軸が一致したフュージョンセンサから得られる画像と Depth 測定値の視差がない情報を得ることが可能である．画像および Depth の解像度は 240×48，フレームレートは 300fps と設定した．カメラと LiDAR は固定した状態で，車やバイク，自転車，人が動く状況をシミュレーションし，200 シーンを作成した．各シーンは 5 秒で，画像と Depth はそれぞれ 300,000 枚から構成される．

学習データのフレームレートおよび空間解像度は，カメラは 6fps で 240×48，LiDAR は 3fps で 120×24 と設定する．CG データの時間分解は 3msec であるため，LiDAR の 1 フレームは 100 枚で 1 セットとなり，Voxel としては 12×2×100 で表される．学習データの総数は 3,000 セットとなる．

3.4 ネットワーク構成および損失関数

ネットワーク構成は UNet[23]の構造を模した全層畳み込みネットワークを用いる．ネットワークの概要を図 1 に示す．

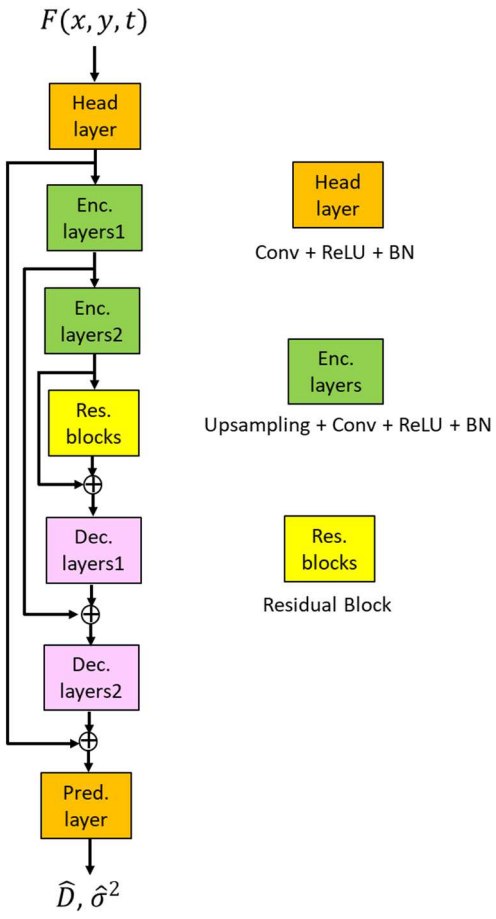

図 1　ネットワークの概要

ネットワークはヘッド層（Head layer），エンコーダー層（Enc. layer），残差ブロック（Res. layer），デコーダー層（Dec. layer），出力層（Pred. layer）で構成し，各エンコーダーとデコーダー層をスキップ接続[24]する．実験ではエンコーダー層は 4 層，残差ブロックは 2 ブロック，ヘッド層の出力チャネル数は 32 と設定する．[11]にならって，ネットワーク $f^{\hat{W}}$ は式（1）の通り Depth 推定値 $\hat{D}$ と不確実性 $\hat{\sigma}^2$ を出力する．

$$\left[\hat{D}, \hat{\sigma}^2\right] = f^{\hat{W}}(F) \tag{1}$$

損失関数は式（2）を用いてネットワークを学習する．

$$L(\theta) = \frac{1}{N}\sum_i \frac{1}{2}\exp\left(\log \hat{\sigma}_i^2\right)\|y_i - \hat{y}_i\| + \frac{1}{2}\log \hat{\sigma}_i^2 \tag{2}$$

式（2）でモデリングした不確実性は，推定が難しいデータの程度を示す Aleatoric Uncertainty[12]を表しており，これを用いて学習することにより，空間的には遠方や物体の境界付近，時間的には目的の時刻に遠い時刻や移動物体のように Depth 推定が困難な Depth 測定値に対して頑健な出力となることを狙う．

3.5 学習条件

学習データの総数は 3,000 セットのうち 2,700 セットを学習用，300 セットをテスト用に用いる．ネットワークは PyTorch[25]を用いて実装した．最適化アルゴリズムは ADAM[26]を用い，学習率は 0.00005，バッチサイズは 1，エポック数は 16 と設定した．

4. 評価

定量的，定性的に我々の超解像の提案手法（以下，Voxel＋マスク＋不確実性）を評価する．また，1 フレーム分の累積（以下，累積），時刻情報を加えた Voxel（以下，Voxel），Voxel にマスクを加えた場合（以下，Voxel＋マスク），累積に不確実性を考慮した場合（累積＋不確実性），Voxel に不確実性を考慮した場合（Voxel＋不確実性）について比較を行う．評価指標は平均平方 2 乗誤差（RMSE）を用いる．

4.1 結果

定量的評価を図 2，3，4 に示す．

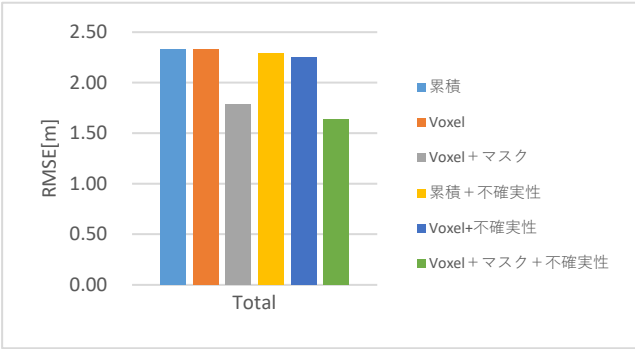

図 2　画面全体の RMSE

図 2 は画面全体の RMSE である．累積と Voxel を比較するとほぼ同様の精度であることが分かる．これらに対して Voxel＋マスク，Voxel＋マスク＋不確実性はいずれも精度

が改善しており，提案手法がもっともいい結果であることが分かる．また，累積，Voxel，Voxel＋マスクの 3 つの手法に対して，それぞれ不確実性を加えるといずれも性能が改善している．

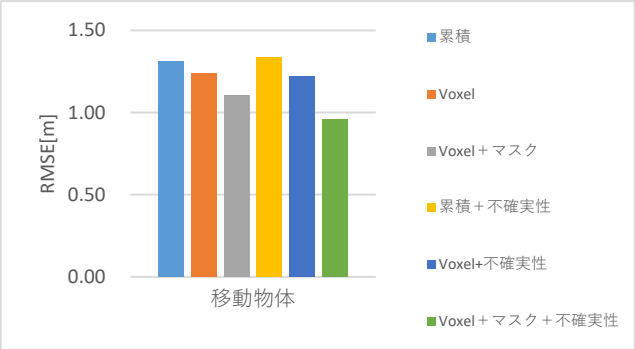

図 3　移動体の RMSE

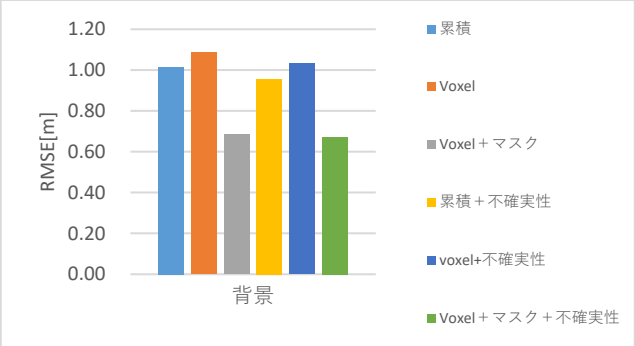

図 4　背景の RMSE

図 3，4 はそれぞれ移動体，背景の RMSE である．累積と Voxel と比較すると，Voxel は累積よりも移動体がよくなっているが，背景は悪化している．これは Voxel では時刻情報を用いることにより移動体の軌跡を類推することができる一方，背景は累積よりも Depth 測定値が疎であるためであると考えられる．そこで Voxel にマスクを加えて背景を高密度化した Voxel＋マスクに注目すると，移動物体と背景のいずれも累積と Voxel よりも性能改善していることが分かる．さらに不確実性を加えた Voxel＋マスク＋不確実性は，Voxel＋マスクよりも移動体の RMSE を改善することができている．損失関数に不確実性を加えたことで移動体に注目して学習していると考えられる．

図 5，6 に定性評価を示す．

図 5 は GT の Depth および累積，Voxel，Voxel＋マスク，Voxel＋マスク＋不確実性を入力とした場合のそれぞれの Depth 推定結果である．赤は近方，緑に近づくにつれ遠方を表している．画面中央の遠方のポールと画面右側の自転車に注目する．GT に対して，累積や Voxel は自転車のエッジが不明瞭であることが分かる．また，Voxel のみでは累積よりもポールのエッジが不明瞭で，図 4 に示す背景の RMSE が Voxel のみでは精度が悪い結果と同じ傾向であることが分かる．Voxel＋マスクおよび Voxel＋マスク＋不確実性に注目すると，Voxel よりもポールのような遠方のエッジが若干改善している．また，自転車のような複雑な形状の Depth が再現されていることが分かる．

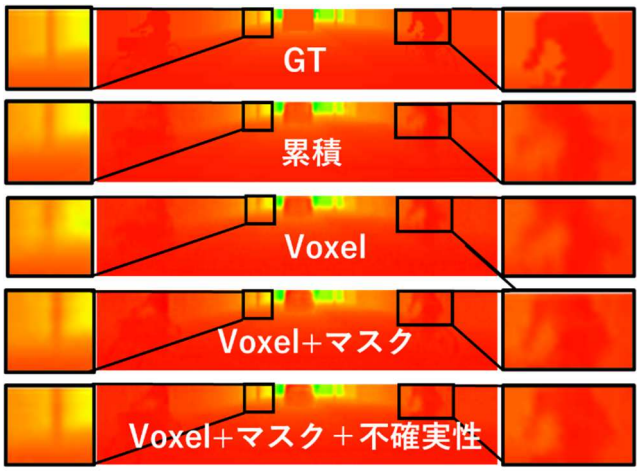

図 5　Depth 推定結果

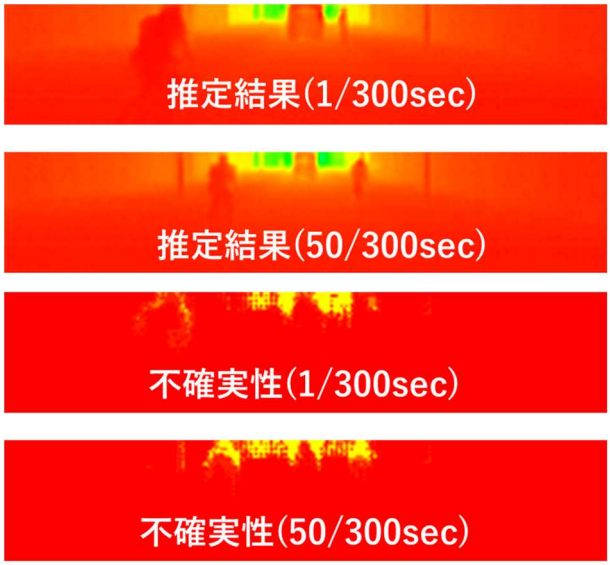

図 6　不確実性

図 6 は不確実性を可視化したもので，移動物体が分かるよう時刻 1/300sec と 50/300sec に注目する．参考に不確実性と併せて推定結果を示す．推定結果については図 5 と同様に赤は近方，緑に近づくにつれ遠方を表している．不確実性については赤が小さく，黄色は大きいことを表している．遠方を除く背景および移動体の中央付近では不確実性が小さいのに対して，遠方の背景および移動体のエッジ付近で不確実性が大きくなっている．この結果から，本研究での狙い通り Depth 推定がそもそも難しい遠方や移動体のエッジ付近に対しては不確実性が大きすぎる場合には損失関数が小さくなり無視できるようにすることにより，Depth 推定の超解像に関する精度について最適化が進む効果が出

ていると考えられる．また，図 3 および図 6 の結果から，学習時に不確実性を考慮することにより移動体，特にその中央付近の精度が向上したことが分かる．

5. まとめ

走査型 LiDAR においてレーザー光のスキャンによる時刻ずれを解消し，移動体に対してずれのない正確な Depth を得るための超解像技術について提案を行った．そのための手法として，走査型 LiDAR で得られた Depth 測定値を時間方向に 3 次元に拡張した Voxel 表現を用いて入力データとして全層畳み込みネットワークで推定する．また，畳み込みが有効に特徴を把握できるようにするため，画像から背景を抽出するマスクを作成することで時空間上の Depth 測定値を高密度化する．さらに，ネットワークの出力として不確実性をモデリングすることで移動体の Depth 推定の精度を向上する．自作の CG データセットを用いた実験により，累積や Voxel を入力とした場合よりも移動体，背景に対して提案手法が有効であることを示した．

参考文献

[1] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2):295–307, 2016.
[2] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in IEEE International Conference on Computer Vision and Pattern Recognition, 1646–1654, 2016.
[3] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in IEEE International Conference on Computer Vision and Pattern Recognition, 2016.
[4] M. S. Sajjadi, R. Vemulapalli, and M. Brown, “Frame-recurrent video super-resolution,” in IEEE International Conference on Computer Vision and Pattern Recognition, 6626–6634, 2018.
[5] Y. Hitomi, J. Gu. Gupta, T., Mitsunaga, and S. Nayar, “Video from a single coded exposure photograph using a learned over-complete dictionary,” in IEEE International Conference on Computer Vision, 287–294, 2011.
[6] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, “High speed and high dynamic range video with an event camera,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.
[7] D. J. C. MacKay, “A practical Bayesian framework for backprop networks,” Neural Computation, 4:448–472, 1992.
[8] A. Graves, “Practical variational inference for neural networks,” in Advances in neural information processing systems, 2348–2356, 2011.
[9] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in Advances in Neural Information Processing Systems, 6402–6413, 2017.
[10] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in International Conference on Machine Learning, pages 1050–1059, 2016.
[11] D. A. Nix and A. S. Weigend. “Estimating the mean and variance of the target probability distribution,” in Proceedings of 1994 IEEE International Conference on Neural Networks, 1:55–60, 1994.
[12] M. Carvalho, B. L. Saux, P. T. Peloux, A. Almansa, and F. Champagnat, “Deep depth from defocus: how can defocus blur improve 3d estimation using dense neural networks?,” in European Conference on Computer Vision, 307–323, 2018.
[13] G. Ros, L. Sellart, J. Materzynska, D. V´azquez, and A. L´opez. “The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in IEEE International Conference on Computer Vision and Pattern Recognition, 2016.
[14] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, “Virtual worlds as proxy for multi-object tracking analysis,” in IEEE International Conference on Computer Vision and Pattern Recognition, 2016.
[15] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite,” in IEEE International Conference on Computer Vision and Pattern Recognition, 3354–3361, 2012.
[16] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, “CARLA: An open urban driving simulator,” in Conference on Robot Learning, 2017.
[17] Epic Games. Unreal Engine 4. https://www.unrealengine.com.
[18] VERTechs. AUTO City. https://www.vertechs.jp/products.
[19] D. Bradley, B. Atcheson, I. Ihrke, and W. Heidrich, “Synchronization and rolling shutter compensation for consumer video camera arrays,” in IEEE International Workshop on Projector-Camera Systems, 2009.
[20] J. Gu, Y. Hitomi, T. Mitsunaga, and S. Nayar, “Coded rolling shutter photography: Flexible space-time sampling,” in IEEE International Conference on Computational Photography, 1 –8, 2010.
[21] X. Cheng, Y. Zhong, Y. Dai, P. Ji, and H. Li, “Noise-aware unsupervised deep lidar-stereo fusion,” in IEEE International Conference on Computer Vision and Pattern Recognition, 2019.
[22] A. Lopez-Rodriguez, B. Busam, and K. Mikolajczyk, “Project to adapt: Domain adaptation for depth completion from noisy and sparse sensor data,” in Proceedings of the Asian Conference on Computer Vision, 2020.
[23] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015.
[24] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, “EV-FlowNet: Selfsupervised optical flow estimation for event-based cameras,” in Robotics: Science and Systems, 2018.
[25] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in Conference on Neural Information Processing Systems Workshops, 2017.
[26] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” in International Conference on Learning Representations, 2015.