深層学習を用いた深度補完の幻覚に関する研究

坪内, 夏輝 / TSUBOUCHI, Natsuki

(出版者 / Publisher)
法政大学大学院理工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学研究科編
(巻 / Volume)
65
(開始ページ / Start Page)
1
(終了ページ / End Page)
6
(発行年 / Year)
2024-03-24
(URL)
https://doi.org/10.15002/00030749

深層学習を用いた深度補完の幻覚に関する研究

A STUDY OF HALLUCINATION PHENOMENA IN DEPTH COMPLETION USING DEEP LEARNING

坪内 夏輝
Natsuki TSUBOUCHI
指導教員 山岸昌夫

法政大学大学院理工学研究科応用情報工学専攻修士課程

This study reports the occurrence of hallucination phenomena in the output depth images generated by the SemAttNet which is a deep learning-based depth completion method known for achieving high accuracy in the popular Depth Completion benchmark of the KITTI dataset. This benchmark dataset collects depth information using observation devices equipped in the front of a driving vehicle. Due to observation settings, the upper part of the depth image corresponding to the sky are not observable in some scenes. In these scenes, we confirm that the generated depth images contain unnatural depth information that appears to be hallucination in the upper part. In addition, inspired by IP-Basic which is a depth completion method that does not use deep learning, we propose a method to remove areas affected by hallucination.
Keywords *: KITTI, Depth Completion, Deep Learning, Hallucinations*

1. はじめに

近年の深層学習の進歩は凄まじく，様々な分野で活用されている．その一例として自動車分野にも広がっており，自動運転車は多くの自動車メーカーが開発に力を入れている．また，高齢化社会の日本では，交通事故への懸念から自動運転技術に対する世間からの関心は高い [1].

AI 技術の発展において学習に用いられる大規模なデータセットは，重要な役割を果たしている．自動運転をターゲットとした代表的なデータセットの一つである KITTI データセット [2] は，2012 年に発表された大規模データセットであり，ベンチマークとして提供されている真値データとの差を比較する深度補完 (Depth Completion) は物体検知に応用される研究として活発に研究が行われている．Depth Completion のベンチマーク [4] が公開された 2017 年以降様々なフレームワークが提案されている [5][6][7]．このベンチマークにおいて優秀な結果を残している多くのフレームワークは深層学習によるものが多い．

一方，自然言語処理などに深層学習を応用した際には，度々学習していないことや学習データとは全く異なるような出力をすることが報告されている [8][9]．Joshua らはテキスト生成において要約モデルが入力には存在しない情報を生成してしまうことを，まるで AI が幻覚を見ているとしてハルシネーションと呼んだ [8]．ハルシネーションは誤情報の拡散を引き起こし兼ねないため，深層学習のリスクの一つとして危険視されている．

本研究では，KITTI データセットの Depth Completion ベンチマークにおいて，高い精度を実現する深層学習を用いた深度補完手法として知られている SemAttNet [10] の出力結果にハルシネーションと思しき現象が発生していることを報告する．具体的には，ベンチマークデータの観測環境の理由により，空に対応する深度画像上部の情報が観測できていない場面が多く含まれている．それらの場面において，深度画像上部にハルシネーションと思しき現実には不自然な深度情報が生成されてしまっていることを確認している．また，深層学習を用いない深度補完手法である IP-Basic のアイデアを参考にして，ハルシネーションを起こしている領域を除去する方法も提案している．

2. 準備

2.1 深度補完

深度補完とは疎な深度画像から密な深度画像を作成するタスクであり，その主な応用先として物体検知や自動運転が挙げられる．深度画像とは距離測定に用いられる深度センサーによって取得された 3 次元的情報を付与された画像である．しかし，距離測定センサーは RGB 画像ほどの情報を得ることが難しく，自動運転車から周辺の深度画像を取得する際に LiDAR を用いると，取得できるデータ自体が疎であったり，被写体によって適切に深度情報を取得できないと言った問題があった．そのため，疎な深度マップから密な深度マップを生成することは物体検知を行う上で重要なタスクである．図 1, 2 は本研究で用いた KITTI データセットに含まれる RGB 画像，疎な深度画像の例である．また，図 3 は図 2 の理想深度画像である．

2.2 KITTI Vision Benchmark Suite

KITTI データセットはドイツのカールスルーエの街を車両前方に搭載された 2 台のステレオカメラと

図 1　KITTI データセットに含まれる RGB 画像

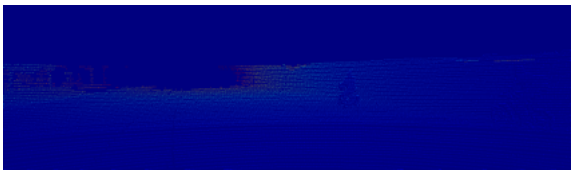

図 2　KITTI データセットに含まれる疎な画像

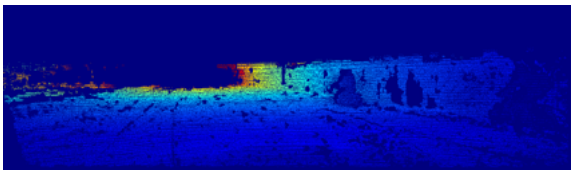

図 3　理想深度画像

LiDAR，GPS 装置によって取得したデータによって構成された，自動運転をターゲットにした大規模なデータセットである [2][3]. データの中には車載カメラによって撮影された未加工のカラー画像とモノクロ画像，LiDAR で撮影された 3D 点群データ，GPS データがある. また，ステレオカメラと LiDAR や GPS 機器の位置を調整するためのキャリブレーションデータと各画像のオブジェクトラベルのデータも含まれている．データには様々な注釈がされており，研究目的に応じた情報を取得することができるため，多くの研究で用いられてきた. また，研究で作成されたデータをもとにした数多くの問題に対するベンチマークが存在する.

KITTI の深度補完のベンチマークである Depth Completion データセットは，深度補完と単一画像深度予測のタスクに対する複雑な深層学習モデルの訓練を可能にする大量の学習データを提供している．このデータセットは，対応する未加工の LiDaR スキャンと RGB 画像を含む 93,000 以上の深度マップからなる．具体的には学習用に 85,000 の疎な深度マップとそれに対応する RGB 画像，検証用に 7,000，テスト用に 1,000 の疎な深度画像によって構成されている．真値は，LiDAR とステレオ推定を累積して生成されている [4].

2.3　SemAttNet による密な深度画像の生成

SemAttNet は Nazir らによって提案された深層学習を用いた深度補完手法 [10] である．従来の深度補完は RGB 画像と深度画像から学習を行なっていた．しかし，それだけでは学習の際に RGB 画像の影など明暗が急激に変化するようなシーンの細部の補完精度が不十分であった．そこで SemAttNet では RGB 画像と深度マップに加え，RGB 画像に写っている物体の意味情報であるセマンティック情報を学習の際に加えることでシーン細部の再現に成功している.

具体的には図 4 のように「主に色情報を用いて深度マップと信頼マップを生成するネットワーク」,「主に画像に映る物体の意味的な情報を用いて深度マップと信頼マップを生成するネットワーク」,「主に深度に関する情報を用いて深度マップと信頼マップを生成するネットワーク」の 3 つのネットワークと，それぞれのネットワークからの出力深度マップ D_{cg}, D_{sg}, D_{dg} と信頼マップ $e^{C_{cg}}$,$e^{C_{eg}}$,$e^{C_{dg}}$ を

$$D_f = \frac{e^{C_{cg}} \cdot D_{cg} + e_{C_{sg}} \cdot D_{sg} + e^{C_{dg}} \cdot D_{dg}}{e^{C_{cg}} + e^{C_{sg}} + e^{C_{dg}}} \quad (1)$$

により統合し，さらに精度を高めることを目的とし既存手法（Atrous 畳み込み [13] を備えた CSPN++[13],[14]）を活用して出力を生成する部分からなる．図 5 に SemAttNet によって生成された深度補完画像の例を示す.

2.4　IP-Basic

IP-Basic は Ku らによって提案された深度補完アーキテクチャである [15]. Ku らはこの研究を通して古典的な画像処理アルゴリズムが深層学習にも匹敵するということを主張している．IP-Basic は，シンプルで高速かつ，CPU 上で実行可能であり，基本的な画像処理操作のみを使用している．さらに IP-Basic では，他の深度補完アーキテクチャとは違い，入力に RGB 画像を必要としない. 以下に IP-Basic の具体的な手順を解説する.

手順 1：KITTI データセットに含まれる深度画像の深度は $2.5m \sim 80m$ の範囲で構成されている．ただし，LiDAR の検知範囲外や鏡面などによって正常に深度が取得できなかった部分などは無効なピクセルとなり，深度値は 0m を取る．そのため，元の深度マップに直接深度補完操作を適用することはできない．そこで，有効なピクセルの深さを反転させバッファを導入する. 反転した深度 D_{inverted} は以下の (2) で表せる.

$$D_{\text{inverted}} = 100.0 - D_{\text{input}} \quad (2)$$

D_{input} は入力深度である．(2) により，有効なピクセルと無効なピクセルの間に 20m のバッファが作成される.

手順 2：有効なピクセルに最も近い無効なピクセルを埋める．無効なピクセルは近くの有効なピクセルと同じ深度値を共有する可能性が高いと考えられる．そこで，有効なピクセルの膨張用に図 6 のような 5×5 のカーネルを設計する．カーネルの形状は，同じ値を持つ可能性が最も高いピクセルが同じ値に膨張されるように設計されている．これにより得られる深度を D_2 とする.

手順 3：手順 2 では全ての無効なピクセルを埋めることができないため，深度マップ内の小さな穴を埋めるための処理をする．図 7 に示す形をした 5×5 のカーネルによって，物体の持つ自然な構造を維持しつつ，小さな穴を閉じる．これにより，近くの深度値が接続され，オブジェクトのエッジが形成される．これにより得られる深度を D_3 とする.

手順 4：手順 3 の操作では中程度の穴は埋まらないため，それらの穴を埋めるために無効なピクセルのマスク（位置情報）を計算し，その後，全てのピクセルに対して図 7 に示す形をした 7×7 カーネルを用いた拡張操作を行う．この操作により，無効なピクセルのみが埋められ,

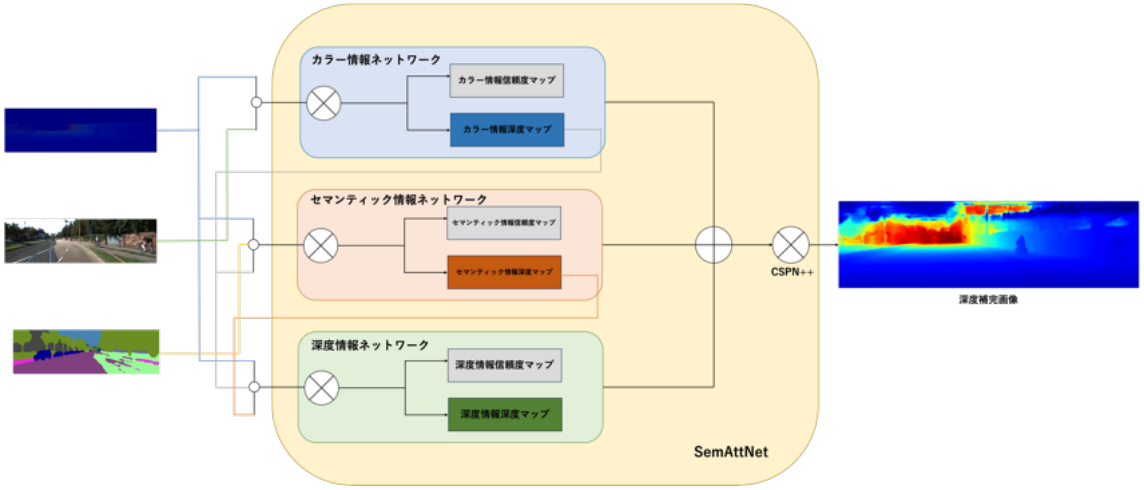

図 4　SemAttNet のネットワーク構造

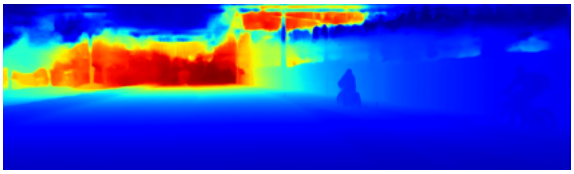

図 5　SemAttNet による深度補完画像

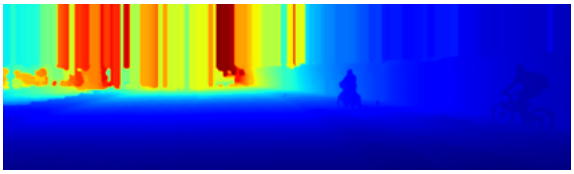

図 8　IP-Basic による深度補完画像

0	0	1	0	0
0	1	1	1	0
1	1	1	1	1
0	1	1	1	0
0	0	1	0	0

図 6　IP-Basic で使用したカーネル形状 1

1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

図 7　IP-Basic で使用したカーネル形状 2

既に計算された有効なピクセルはそのまま保持される．これにより得られる深度を D_4 とする．

手順 5：ここまでの操作では LiDAR による検知範囲外にある無効なピクセルががそのまま残っているので，ピクセルの各列の最上部の深度をとり，上部の空ピクセルに適用する．これにより無効なピクセルのない深度画像が生成できる．こららの操作でまだ残っている無効なピクセルのある大きな穴は図 7 と同等の形状をした 31×31 のカーネルを使用することで近くの深度値から値が挿入される．これにより得られる深度を D_5 とする．

手順 6：以上の手順で密な深度マップは得られるのだが，深度補完を行う際に発生した外れ値を取り除くため，5×5 のガウシアンブラーをかける．これにより，全体を滑らかにし，エッジを取り除く．最後に初めの段階で反転させていたため，ピクセルを反転させて元に戻すことで以下の最終的な深度を得る

$$D_{\text{output}} = 100.0 - D_5 \tag{3}$$

以上によって得られた深度画像を図 8 に示す．

3.　深層学習を用いた深度補完手法 SemAttNet の幻覚とその検出

3.1　SemAttNet の幻覚

SemAttNet によってハルシネーションが生じていることを確認するため，KITTI データセットの疎な深度画像と RGB 画像から SemAttNet を用いて密な深度画像の生成を行なった．SemAttNet にはセマンティック画像も併せて入力する必要があるが，KITTI データセットには RGB 画像に紐付けられたセマンティック情報が提供されていない．そこで，[10] に倣い，WideResNet38 [16] を使用して RGB 画像からセマンティック情報を得た．図 9 は，RGB 画像から WideResNet38 を用いて生成されたセマンティック画像である．

図 10，11 は SemAttNet によって生成した密な深度画像を 3 次元表示したものである．場面は図 1 の RGB 画像と同じである．図 11 と図 12 を見ると，SemAttNet による深度補完画像は真値と比べて上側に深度が追加されていることがわかる．しかし，図 1 に示す対応す

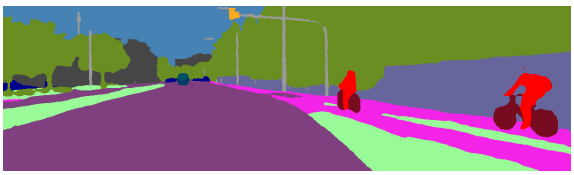

図 9　WideResNet38 による予測セマンティック画像

る RGB 画像には上側の部分には空や奥にある茂みしか写っていない．図 11 では，手前にいる自転車に乗った男性のシルエットよりも RGB 画像における茂みの部分の深度の方がかなり手前にあることがわかる．また，深度情報を取得するために使用した車載レーザースキャナの LiDAR は水平視野角は 360 度であるのに対し，垂直視野角は 26.9 度である．これは例えば 5m 先の物体は約 2.5m を超える高さだと，超えた分の深度を取得することができないということである．そのため，LiDAR の検知視野角を超えた部分に深度が発生することは考えられず，図 10 の茂みの部分や空の深度は本来取得されるはずのない深度である．

SemAttNet によって生成された密な深度画像は入力の疎な深度画像に存在する有効なピクセルの領域では精度の高い深度補完がされているが，入力に存在しない領域，すなわち LiDAR の検知範囲外では，精度の低い深度が追加されている．これは，深度補完を行う際に，疎な深度マップ以外に RGB 画像やセマンティック情報が入力されているため，疎な深度画像に存在せず，その他の入力に存在している領域の情報から予測深度を生成しているためであると考えられる．Depth Completion が物体検知や自動運転につながるタスクだとするならば，本来近い距離に存在しない深度を発生させているこの現象は看過できないものである．

本研究ではこのような，深層学習で生成した深度画像において，LiDAR の検知範囲外にであるために入力の疎な深度画像や真値画像には存在しない領域に誤った深度を持つ部分を深層学習によるハルシネーションとした．また，LiDAR の検知範囲外や物体の表面の性質などにより適切に深度が取得できておらず，深度値が存在しないまたは，深度が 0m になっているピクセルを無効なピクセルとし，検知可能距離範囲内で検知が適切に行われている，若しくは深度補完によってその距離範囲内の深度値を持つピクセルを有効ピクセルとした．

3.2　IP-Basic を用いた有効深度の境界の特定

SemAttNet によるハルシネーションを検出するために，IP-Basic による深度補完のアルゴリズムからハルシネーションであると推測される領域と入力のそな深度画像に真値が存在し，適切に学習および深度補完が行われている領域との境界を特定する．具体的には IP-Basic の手順 1〜手順 5 を用いることで，有効深度と無効なピクセルの画像上部との境界を特定することを提案する．IP-Basic は，疎な深度画像に対して忠実に深度補完を行うため，この境界が LiDAR のデータ取得範囲の上端に近い値を持つことが期待できる．また，SemAttNet によって生成された深度画像において，境界から上部を無

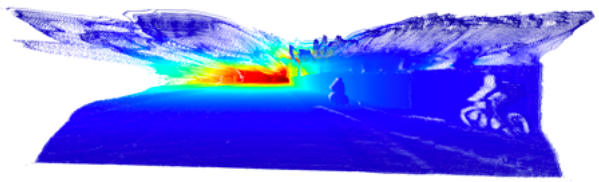

図 10　3 次元表示した SemAttNet による密な深度画像 (正面)

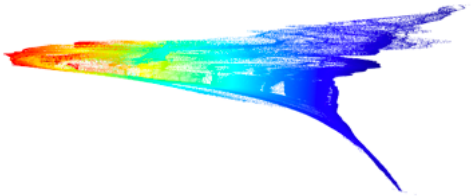

図 11　3 次元表示した SemAttNet による密な深度画像 (横)

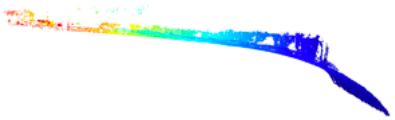

図 12　3 次元表示した真値深度画像 (横)

効なピクセルで置き換えることにより，ハルシネーションの大部分の除去が実現できると考えられる．

4.　数値実験：ハルシネーションの除去

本実験では，SemAttNet によって出力された 1000 枚の深度補完画像に対し，ハルシネーションの除去を行なった．その後，深層学習による深度補完画像から，ハルシネーションの部分を取り除いた深度画像と真値とのスコアを比較する．表 1 には SemAttNet，IP-Basic，そして 3.2 節の手法で作成した深度画像の RMSE(平方平均誤差)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(D_{\mathrm{gt},i} - D_{\mathrm{pred},i})^2} \qquad (4)$$

のスコアを示す．ここで，D_{gt} は真値の深度，D_{pred} は予測深度，N は真値の有効ピクセル数を表す．

表 1　各方法における深度補完画像の評価

方法	RMSE(mm)
SemAttNet	738.13
IP-Basic	1350.93
幻覚を取り除いた深度画像	741.04

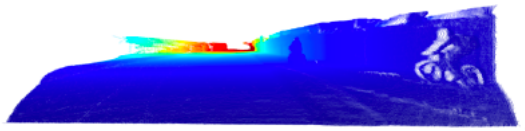

図 13　ハルシネーションを除去した深層学習による深度補完画像 (正面)

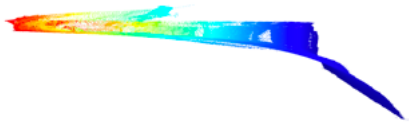

図 14　ハルシネーションを除去した深層学習による深度補完画像 (横)

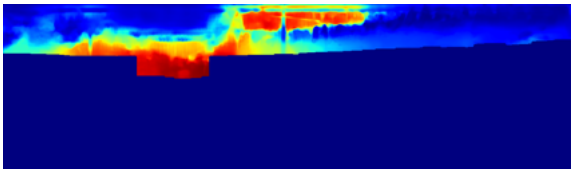

図 15　提案手法によりハルシネーションと断定した部分

図 13，14 にはハルシネーションを取り除いた深層学習による深度画像を 3 次元表示したものを示す．また，図 15 に 3.2 節の手法によって切り取られた密な深度画像のハルシネーション部分のみの画像を示す．

SemAttNet によるほぼ全てのシーンの深度画像では有効な深度を除去することなく，ハルシネーションを除去する前の画像から RMSE スコアはほぼ変化がなかった．しかし，深度画像からハルシネーション部分を除去した深度画像が除去前の本来の深度補完画像に比べ特に RMSE のスコアが悪化しているシーンが 2 つ見つかった．これらのシーンでは，ハルシネーションを除去した際に有効な深度も同時に除去してしまっており，それがスコアの悪化の原因となっていると考えられる．図 16，17 に特に劣化が見られる深度画像を示す．

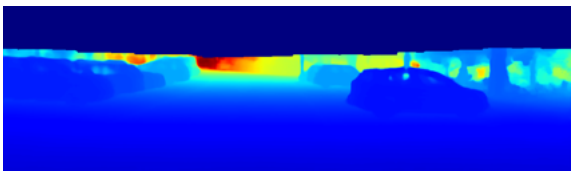

図 16　RMSE の劣化が大きい幻覚除去画像 1

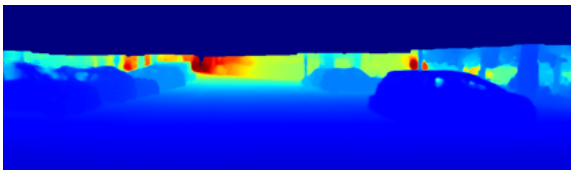

図 17　RMSE の劣化が大きい幻覚除去画像 2

図 18　真値とハルシネーション除去した深度補完画像の有効ピクセルの差の例

図 19　図 18 に対応した RGB 画像

これら 2 つの場面の深度画像に共通するのは真値が入力の疎な深度画像と比較して全体的に有効深度の描写範囲が上にずれているためである．図 20 は図 16 に対応した入力の疎な深度画像と真値の深度画像を並べたものに横線を入れたものである．図 20 を見ると，疎な深度画像には上から 2 本目のグリッド線より上に深度描写がないのに対し，真値には描写されており，疎な深度画像と真値で有効深度の範囲がずれていることがわかる．これは，真値を作成する際に LiDAR による深度マップを RGB 画像をもとに位置合わせを行なっているためである．

また，図 18 は図 16 の深度画像において真値に存在する有効ピクセルの中で，ハルシネーション除去を行った際に誤って除去されてしまった範囲を赤で強調表示した画像である．図 19 は図 18 に対応した RGB 画像である．入力に用いられている疎な深度画像は LiDAR で取得した疎な深度マップをそのまま使用している．そのため，入力の疎な深度画像に対して忠実に深度補完を行った IP-Basic の深度補完画像は位置合わせによる真値のズレに対応していない．しかし，SemAttNet は入力進度と RGB 画像の位置合わせを行なっているため，真値とそな進度画像のずれに対応できたと考えられる．これは疎な深度マップから密な深度マップを作成する深度補完というタスク行う上で RGB 画像を用いた位置合わせを

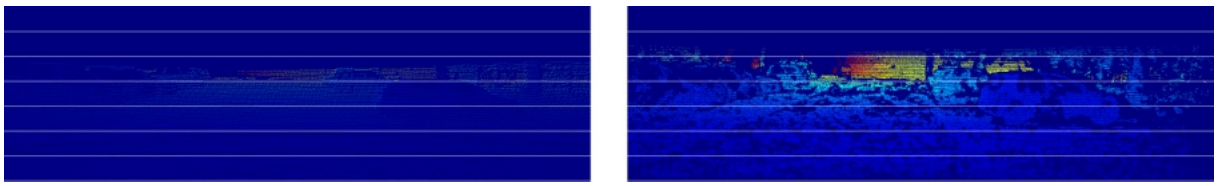

図 20　図 19 に対応した入力の疎な深度画像 (左) と真値の深度画像 (右) に白線を入れた画像

行わなければ優秀なスコアを出すことができないことを表している．

5. まとめ

本研究では，KITTI データセットを用いて，深層学習を用いた深度補完手法 SemAttNet がハルシネーションを引き起こすことがあることを明らかにした．また，KITTI データセットの特徴と IP-Basic の手続きを組み合わせることで，ハルシネーションの発生箇所を特定する手法についても提案した．また，提案手法によりハルシネーションを除去した深度画像と真値の比較を行い，良好な結果を得た．今後の課題として，ハルシネーションの発生そのものの抑制や，深層学習手法の改善，より正確なセマンティック情報の利用などに焦点を当て，深度補完手法の信頼性の向上に繋げたい．

参考文献

[1] 宮木 由貴子, “社会的受容性の醸成に向けた調査と評価,” *SIP* **成果報告書**, vol. 2022, no. 1, pp. 165–171, 2022.

[2] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITII vision benchmark suite,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[3] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITII dataset,” *International Journal of Robotics Research (IJRR)*, 2013.

[4] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, “Sparsity invariant CNNs,” in *International Conference on 3D Vision (3DV)*, 2017.

[5] Z. Yan, K. Wang, X. Li, Z. Zhang, J. Li, and J. Yang, “RigNet: Repetitive Image Guided Network for Depth Completion,” in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 214–230.

[6] Y. Wang, B. Li, G. Zhang, Q. Liu, T. Gao, and Y. Dai, “Lrru: Long-short range recurrent updating networks for depth completion,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 9422–9432.

[7] S. Shao, Z. Pei, W. Chen, X. Wu, and Z. Li, “ND-Depth: Normal-distance assisted monocular depth estimation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7931–7940.

[8] Joshua Maynez and Shashi Narayan and Bernd Bohnet and Ryan Thomas Mcdonald, “On Faithfulness and Factuality in Abstractive Summarization,” in *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[9] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[10] D. Nazir, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, “SemAttNet: Towards attention-based semantic aware guided depth completion,” *IEEE Access*, pp.120781–120791, 2022.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[12] F. Fooladgar and S. Kasaei, “Multi-Modal Attention-based Fusion Model for Semantic Segmentation of RGB-Depth Images,” *CoRR*, vol. abs/1912.11691, 2019.

[13] X. Cheng, P. Wang, C. Guan, and R. Yang, “CSPN++: learning context and resource aware convolutional spatial propagation networks for depth completion,” Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 7, 2020.

[14] M. Hu, S. Wang, B. Li, S. Ning, L. Fan, and X. Gong, “PENet: Towards precise and efficient image guided depth completion,” IEEE International Conference on Robotics and Automation (ICRA), pp. 13656-13662, 2021.

[15] J. Ku, A. Harakeh, and S. L. Waslander, “In defense of classical image processing: Fast depth completion on the CPU,” IEEE Conference on Computer and Robot Vision (CRV), 2018.

[16] Z. Wu, C. Shen, and A. van den Hengel, “Wider or Deeper: Revisiting the ResNet Model for Visual Recognition,” *Pattern Recognition*, vol. 90, pp. 119–133, 2019.