Vol.2020-CVIM-222 No.19
2020/5/15

ワイヤフレームとの同時学習による単一画像からの深度推定

水沼 佑太[1,†1,a)]　数藤 恭子[1,b)]

概要：近年，単一画像からの機械学習を用いた深度推定手法が多数試みられている．しかし，十分な精度を得ることは難しく，その一つの要因として物理的な情報が不足していることが考えられる．そこで本研究では三次元構造の情報を補うデータを同時に学習する機械学習ベースの手法を提案する．具体的には，深度とワイヤフレームを同時に学習するネットワークを構築する．ワイヤフレームは，画像上で消失線とみなせるような平行な線分の組を含むことから深度情報を補う可能性と，また，平面同士の境界線を構成することから境界付近の推定精度を高める可能性を期待した．評価実験では，構築したネットワークを用いて深度のみを学習した場合と比較した．目視による評価では，一部の入力画像において，提案手法により線分で構成される人工物の輪郭や壁と床の境界線付近などで物理的な矛盾の減少が確認された．また，画素値の誤差を表す RMSE や，画像の類似度を表す SSIM を用いた評価では，提案手法により局所的に良好な学習ができることが確認された．

Depth Estimation from Single Image by Simultaneous Learning with Wireframe

Abstract: In recent years, many depth estimation methods using machine learning from a single image have been attempted. However, if the information on the three-dimensional structure of the training data is insufficient, inconsistency may occur. In this article, we propose a machine learning-based method that simultaneously learns data that supplements three-dimensional structure information. Specifically, we construct a network that learns depth and wireframe at the same time by using wireframe as 3D structure information. Wireframes include a set of parallel line segments that can be regarded as vanishing lines on the image, We expected the possibility that supplementing depth information is improved, and the accuracy of estimation near the boundary is improved by forming a boundary between planes. In the evaluation experiment, we compared it with the case where only the depth was learned using the constructed network. In the visual evaluation, in some input images, it was confirmed that the proposed method reduced the inconsistency in the contour of the artificial object composed of line segments, and near the boundary between the wall and the floor. In the evaluation using RMSE, which represents the error of the pixel value, and SSIM, which represents the similarity of the images, it was confirmed that the proposed method could locally perform good learning.

1. はじめに

単一画像からの深度推定は，自動運転技術や医療分野などで深度情報を取得する際に，センサを利用することが難しい場合や既に撮影された画像しかない場合などにおいて活用が期待されている．本来，画像からの深度推定は視差のある複数枚の画像が必要であるが，単一画像からは物体の陰影やオクルージョンの状態から三次元構造を推定することができる．近年はこれと同様の推定を機械学習により学習したモデルに行わせる試みも多数示されている．その一つのアプローチは，エンコーダデコーダ型のネットワークで深度画像を生成するもので，一度モデルを構築すれば単一画像から簡易に深度画像を生成できることが利点である．しかしこの手法では，画像を再構成する際に情報が失われるため，特に人工物の輪郭や面の境界付近において粗くなりがちなことが問題である．その結果，図 1 に示すように，しばしば物理的な矛盾（人工物の輪郭や壁面などでの歪み）が生じる．

[1] 東邦大学 〒 274–8510 千葉県船橋市三山 2–2–1
Toho University Miyama 2–2–1, Funabashi-shi, Chiba 274–8521, Japan
[†1] 現在，筑波大学
Presently with University of Tsukuba
[a)] mizunuma@le.iit.tsukuba.ac.jp
[b)] kyoko.sudo@sci.toho-u.ac.jp

(左)RGB 画像，(中央) 正解画像，(右) 生成画像

図 1 物理的な矛盾が生じた例
Fig. 1 Example of inconsistency

図 2 深度推定のイメージ
Fig. 2 Image of depth estimation

そこで本研究では，機械学習をベースとした単一画像からの深度推定において，こうした不具合を低減するために有効と考えられる新たな付加的な情報を同時に学習することを提案する．なお本稿では，深度推定は入力の RGB 画像の各画素に対して推定した深度をグレースケール画像として出力することと定義する．深度推定のイメージを図 2 に示す．左から順に RGB 画像，深度画像（各画素が持つ深度をグレースケールで表した画像）である．

2. 関連研究

深度推定は視差のある複数枚の画像から被写体の形状とカメラの位置を推定する SfM[1] や，レーザなどのセンサが取得した情報から自己位置推定と地図作成を同時にする SLAM[2] による手法などが取られてきた．しかし，シーンによっては複数枚の画像を用意するのが難しく，十分な情報を得られない．近年はその不足する情報を学習によって補うことで深度を推定する，機械学習ベースの研究が盛んに行われている．

Karsch らは類似したシーンの画像は類似した深度値を持つとして，入力画像に類似する画像の候補を RGBD 画像のデータベースから見つけ，調整と補完をすることで深度を推定した [3]．佐藤らは対となっている画像のペアで構成される学習データの関係を学習し，入力画像を対応する画像に変換する画像生成アルゴリズム [4] を用いて入力画像の各ピクセルに対して推定した深度をグレースケール画像として出力することで深度を推定した [5]．これらの手法では生成画像が正解画像と類似した画像となるように学習される．本研究では画像中のオブジェクトから得られる情報を学習に用いることで推定精度の向上を試みる．

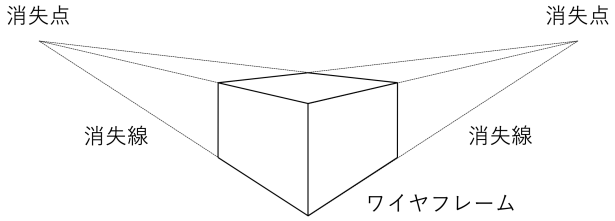


図 3 ワイヤフレームと消失線の関係のイメージ図
Fig. 3 Image of the relationship between wireframes and vanishing lines

3. 提案手法

3.1 提案手法の概要

人工物の輪郭や面の境界付近での推定精度を向上させるための付加情報として，ワイヤフレーム（人工物の輪郭）の利用が考えられる．ワイヤフレームは，画像上で消失線とみなせるような平行な線分の組を含むことから（図 3），それらの線分によって構成されるオブジェクトの見かけ上の変化から深度情報が得られる可能性と，また，平面同士の境界線を構成することから境界付近の推定精度を高める可能性が期待される．

本稿では，これを学習する仕組みとして，入力を RGB 画像として深度画像とワイヤフレーム画像（ワイヤフレームのみ描画した画像）を出力するようなネットワークを構築する．ネットワーク構造は，入力と出力が画像のタスクにおいて広く用いられている，エンコーダデコーダ型の畳み込みニューラルネットワークである pix2pix をベースとする．ただし，ワイヤフレーム画像を生成するエンコーダの勾配を用いて深度画像を生成するエンコーダを最適化する構造を取り入れた．これにより，深度画像を生成するエンコーダは深度とワイヤフレームの双方に対して互いに矛盾の少ない特徴（物理的な矛盾の少ない特徴）が得られると期待した．構築したネットワーク構造の全体図を図 4 に示す．Generator は Discriminator が生成画像に対して Real と判定させるように学習し，Discriminator は生成画像に対して Fake，正解画像に対して Real と判定するように学習する．

3.2 データセット

学習には深度とワイヤフレームの情報が必要であるが，それらがペアとなったデータセットを作成するのは困難である．そこで，NYU Depth Dataset V2[7] と ShanghaiTech dataset[8] を用いて深度とワイヤフレームの情報をそれぞれ取得することでそれに替えた．NYU Depth Dataset V2 は Microsoft Kinect の RGB カメラと Depth カメラの両方で記録された様々な屋内シーンにおけるビデオシーケンスである．学習に使用した NYU Depth Dataset V2 の例を図 5 に示す．ShanghaiTech dataset[8] は人工環境における

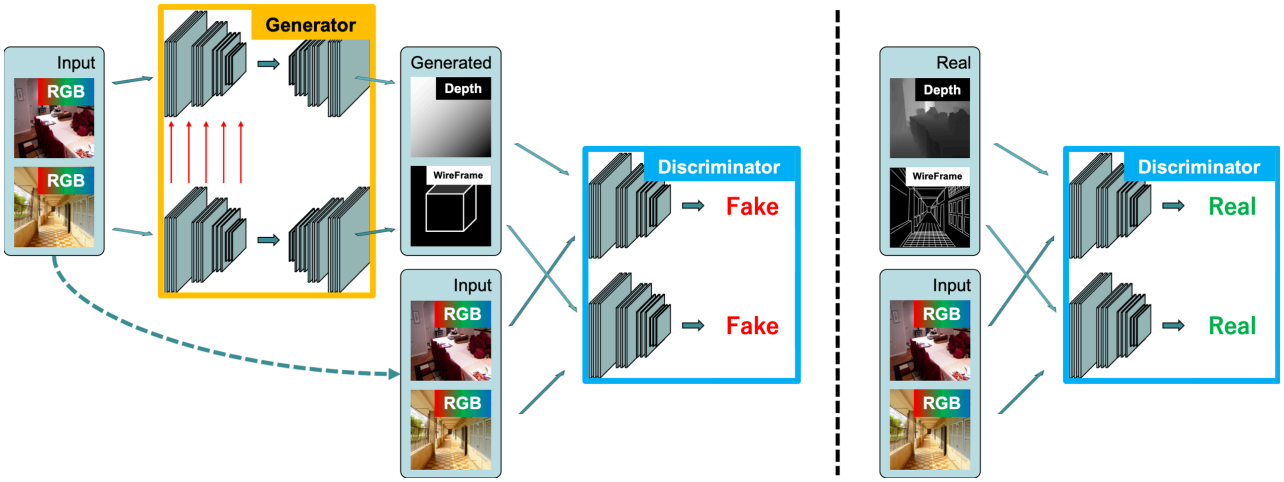
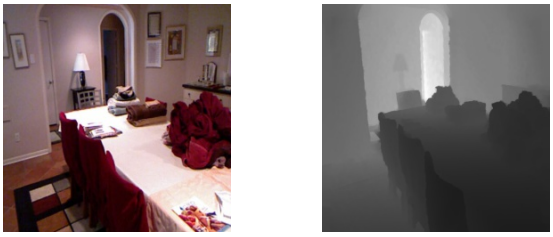
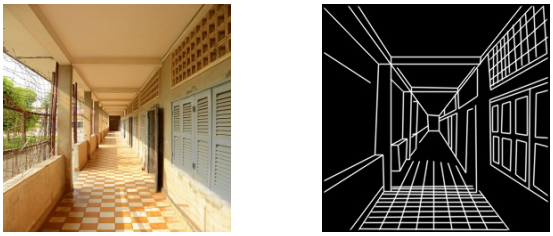


図 4　ネットワーク構造の全体図
Fig. 4　Overview of network structure



RGB 画像　　深度画像

図 5　NYU Depth Dataset V2 の例
Fig. 5　Exapmple of NYU Depth Dataset V2



RGB 画像　　ワイヤフレーム画像

図 6　ShanghaiTech dataset の例
Fig. 6　Example of ShanghaiTech dataset

RGB 画像と，その画像内のワイヤフレームを構成する点や直線などの情報がまとめられたデータセットである．学習に使用した ShanghaiTech dataset の例を図 6 に示す．

3.3　ネットワークの学習過程

次の過程を 1 イテレータ毎に繰り返し，学習した．

- 深度画像を生成するネットワークの学習（3.3.1）
- ワイヤフレーム画像を生成するネットワークの学習（3.3.2）
- ワイヤフレーム画像を生成するエンコーダの勾配を用いて，深度画像を生成するエンコーダを最適化

3.3.1　深度画像を生成するネットワークの学習

(1) Generator に RGB 画像を入力する．
(2) U-Net[6] を経て深度画像が出力される．
(3) Discriminator に RGB 画像と深度画像（生成画像），RGB 画像と深度画像（正解画像）をそれぞれ入力する．
(4) PatchGAN を経て生成画像と正解画像に分類される（特徴マップが出力される）．
(5) 重みを最適化する．

3.3.2　ワイヤフレーム画像を生成するネットワークの学習

(1) Generator に RGB 画像を入力する．
(2) U-Net を経て生成画像が出力される．
(3) Discriminator に RGB 画像とワイヤフレーム画像（生成画像），RGB 画像とワイヤフレーム画像（正解画像）をそれぞれ入力する．
(4) PatchGAN を経て生成画像と正解画像に分類される（特徴マップが出力される）．
(5) 重みを最適化する．

4.　評価実験

本稿では提案手法（深度とワイヤフレームを同時学習する場合）と pix2pix（深度のみを学習する場合）を比較した．具体的には，それぞれの手法によって生成された深度画像に対して，目視による比較と評価指標による比較をした．目視による比較では人工物の輪郭や壁面などで物理的な矛盾の減少が見られるかを確認し，評価指標による比較では生成画像を用いてネットワークの学習精度を確認した．

4.1　実験条件

NYU Depth Dataset V2 から，深度の欠損値が埋められた Labeled Dataset を使用する．Labeled Dataset は 1449 ペアの画像で構成されており，訓練用データを 1304 ペア，テスト用データを 145 ペア用いた．ShanghaiTech dataset は 5000 ペアの訓練用データと 462 ペアのテスト用データで構成されており，訓練用データを NYU Depth Dataset

V2 と同数の 1304 ペア，テスト用データには 462 ペア用いた．なお，RGB 画像と深度画像は [0, 255] に正規化し，256 × 256 にリサイズした．ShanghaiTech dataset はワイヤフレーム情報が座標データで格納されており，OpenCV の線分描画関数でワイヤフレーム部分を 255，背景を 0 として描画し，ガウシアンフィルタを施したものをワイヤフレーム画像とした．以上を用いて，ネットワークに入力する際には画素値は [-1, 1] に正規化し浮動小数点に変換した．また，訓練データにはランダムクロップによりデータ拡張を行った．最適化手法は学習率 2e-4, 減衰率 0.5 の Adam とし，150 エポック学習を行った．

評価指標は RMSE[9]（平均平方二乗誤差）と SSIM（構造的類似性指数）[10] を用いる．RMSE は（1）の式で表し，生成画像と正解画像の画素値の差が小さいほど小さい値を取る．ただし，n は画素数，f_i は生成画像の画素値，y_i は正解画像の画素値である．SSIM は（2）の式で表し，生成画像と正解画像が類似するほど大きい値を取る．ただし，μ_x，μ_y は生成画像と正解画像のそれぞれの局所領域における画素値の平均値，σ_x，σ_y は標準偏差，σ_{xy} は共分散である．また，$c_1 = 0.01$，$c_2 = 0.03$ とした．

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(f_i - y_i)^2} \tag{1}$$

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{({\mu_x}^2 + {\mu_y}^2 + c_1)({\sigma_x}^2 + {\sigma_y}^2 + c_2)} \tag{2}$$

4.2 実験結果

4.2.1 目視による比較

提案手法と pix2pix のそれぞれの生成画像（深度画像）の差異はほとんど見受けられなかったが，提案手法で物理的な矛盾の減少が一部見られた．具体的には，図 7 に示すように，線分で構成される人工物の輪郭や壁と床の境界線付近などで歪みや前後関係の矛盾の減少が見られた．また，図 8 の生成画像（ワイヤフレーム画像）からは，ワイヤフレームの特徴を学習できていることが見受けられた．これらのことより，深度推定においてワイヤフレームとの同時学習が良好な影響を僅かながら与えたと考えられる．

4.2.2 評価指標による比較

RMSE と SSIM を 1 エポック毎に取得し，それぞれ図 9 と図 10 を得た．これらより，提案手法はワイヤフレームを同時に学習する分，pix2pix よりも得られる値にややばらつきが生じる様子が見受けられた．また，これらの値について，表 1 のように RMSE と SSIM の平均値，最終値 (150 エポック)，取得した中で最も良好な値である最小値（RMSE）/最大値（SSIM）を比較し，表 1 を比較したところ，平均値は提案手法が僅かに下回ったものの，差異は少なかった．最終値，最小値（RMSE）/最大値（SSIM）では提案手法で良好な値が得られた．

図 7　物理的な矛盾の減少が見られた生成画像 (深度画像) の例
Fig. 7　Example of generated image (depth image) with reduced inconsistency

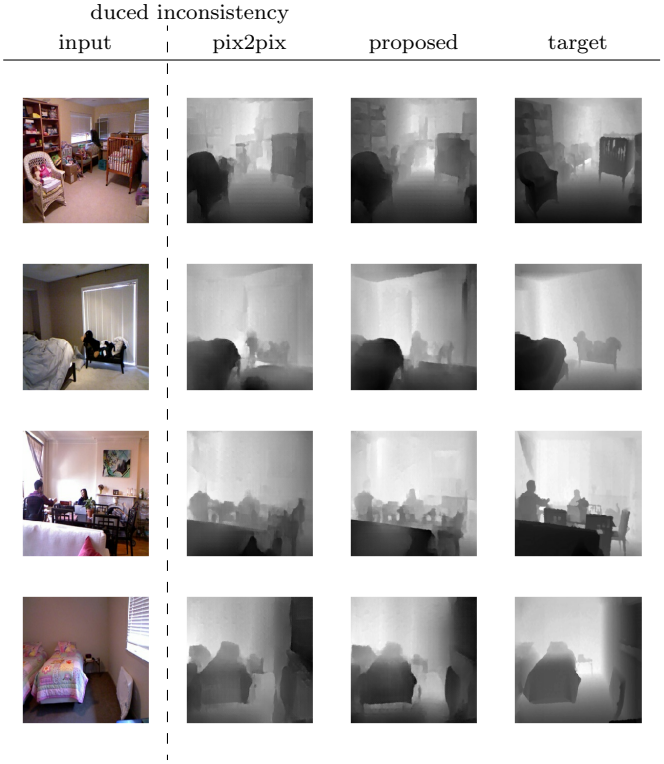


図 8　生成画像 (ワイヤフレーム画像) の例
Fig. 8　Example of generated image (wireframe image)

5. まとめ

提案手法は大幅な精度向上は望めないが，局所的には良好な学習ができるといえる．評価指標による比較において

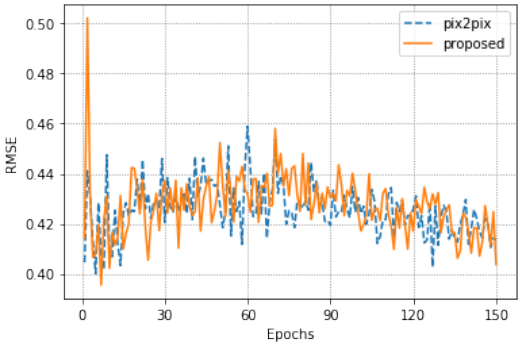


図 9　学習による RMSE の変化
Fig. 9　Transition of RMSE by training

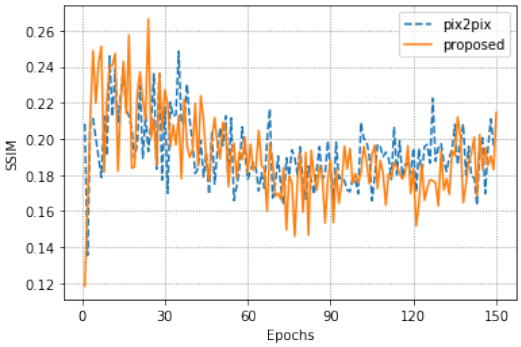


図 10　学習による SSIM の変化
Fig. 10　Transition of SSIM by training

表 1　RMSE と SSIM の平均値，最終値 (150 エポック)，最小値 (RMSE)/最大値 (SSIM)
Table 1　Average value, final value (150 epochs), minimum value (RMSE) / maximum value (SSIM) of RMSE and SSIM

	RMSE			SSIM		
	平均値	最終値	最小値	平均値	最終値	最大値
pix2pix	0.425	0.413	0.400	0.192	0.196	0.248
proposed	0.427	**0.403**	**0.395**	0.191	**0.214**	**0.266**

平均値で提案手法が下回った原因として，深度とワイヤフレームはそれぞれ異なる最適解を持っており最適化が難しくなったと考えられることから，今後は適切な学習率の設定が課題であると考える．また，物体の前後関係や面の境界の明瞭度といった，三次元的な構造の推定の正しさを定量的に評価できる評価指標の検討が必要であると考える．

参考文献

[1] Shimon Ullman: *The Interpretation of Structure from Motion*, Proceedings of the Royal Society of London, vol. 203, no. 1153, pp. 405–426 (1979).
[2] Hugh Durrant-Whyte and Tim Bailey: *Simultaneous localization and mapping: part I*, IEEE Robotics & Automation Magazine, vol. 13, no. 2, pp. 99–110 (2006).
[3] Kevin Karsch, Ce Liu and Sing Bing Kang: *Depth Extraction from Video Using Non-parametric Sampling*, European Conference on Computer Vision, pp. 775–788 (2012).
[4] Phillip Isola et al.: *Image-to-Image Translation with Conditional Adversarial Networks*, 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 5967–5976 (2017).
[5] 佐藤 颯人 et al. : 機械学習による *RGB* 画像からの距離画像の生成, 第 80 回全国大会講演論文集, 2018 巻, 1 号, pp. 229–230 (2018).
[6] Olaf Ronneberger, Philipp Fischer and Thomas Brox: *U-Net: Convolutional Networks for Biomedical Image Segmentation*, Medical Image Computing and Computer-Assisted Intervention, vol. 9351, pp. 234–241 (2015).
[7] Nathan Silberman et al.: *Indoor Segmentation and Support Inference from RGBD Images*, European Conference on Computer Vision, pp. 746–760 (2012).
[8] Kun Huang et al.: *Learning to Parse Wireframes in Images of Man-Made Environments*, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 626–635 (2018).
[9] Rob J. Hyndman and Anne B. Koehler: *Another look at measures of forecast accuracy*, International Journal of Forecasting, vol. 22, no. 4, pp. 679–688 (2006).
[10] Zhou Wang et al.: *Image quality assessment: From error visibility to structural similarity*, IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612 (2004).