CH-001

深度と深度勾配の相互変換による Multi-task learning を活用した教師なし単眼深度推定の提案
Proposal of unsupervised monocular depth estimation using multi-task learning by mutual conversion of depth and depth gradient.

髙嶺 潮† 遠藤 聡志†
Michiru Takamine Satoshi Endo

1 はじめに

深度はカメラから被写体までの距離を表す空間情報の一種であり，その取得方法は大別して三つ存在する．一つ目は LIDAR 等の特殊なセンサーを用いて光学的に深度を測定する手法であり，二つ目は視差画像などから深度を予測する複眼深度推定である．前者は設備費用が高く，後者は観測を複数回必要とする脆弱性を持つ．また，両者は密な深度の取得を不得意とし，欠損部分を保管する技術との併用が不可欠である [1][2]．対して，RGB 画像 1 枚からピクセル単位の深度を予測する単眼深度推定は上記の欠点を克服しており，空間情報を取得する重要な手段の 1 つとして扱われている．過去には，単眼での深度推定は不良設定問題であるため推定が難しいとされてきたが，機械学習技術の進歩によって数々の実用的なモデルが登場するに至った [3]．同時に，正確な教師データ (RGB-D image) の入手が困難であるという根本的な問題が単眼深度推定には残されており，これからは教師データに依存しない教師なし学習手法の開発が求められている [4]．また，教師あり学習においては深度勾配の Multi-task learning ならびに Multiple-input が深度推定に大きく貢献することが判明しているが [5][6]，教師なし学習に最適化した例は少ない．以上を踏まえ，本稿では，深度勾配推定の Multi-task learning に適用した教師なし学習モデルを提案し，単眼深度推定の推定精度向上を目指す．

2 単眼深度推定の動向と考察

単眼深度推定は不良設定問題であるため，経験則や仮定を前提とする知識をモデルに与える工夫が必要となる．この問題に対処すべくヒューリスティクスを活用する研究が最初期に行われたが，それらはいずれも入力画像に大きな制限が課せられていた [7]．Eigen ら [8] は CNN で人間の深度推定を模倣しようと試み，大域深度と局所深度を段階的に求めることで，入力画像に制限を設けない深度推定を可能とした．上記で提案された Multi-Scale Model:MSM とその改善モデル [9] の活躍により，機械学習手法を用いた単眼深度推定におけるベースラインが打ち立てられ，その際の推定精度は Relative absolute error:abs rel(%) において 15.8% を記録した．その後，人間が周辺情報を使ってタスクを補完する方法を参考にした深度推定の研究が幾つか発表されている．Li らは深度勾配を Multiple-input として活用する Two-Streamed Network:TwoNet[6] を提案し，abs rel において 14.3% の推定精度を実現している．単眼深度推定に Multi-Task learning を組み合わせた代表的なモデルとしては Semantic label の推定問題を取り扱った Jafari ら [5] の Joint Refinement Network:JRN と Surface normal の推定問題を取り扱った Qi ら [10] の Geometric Neural Network:GeoNet が存在する．これらの研究によって，モデルに追加情報を与える手法の有用性が示されると同時に，人手で作成されたラベルの曖昧性が学習に悪影響を及ぼす例も確認されている [11][12]．また，教師あり単眼深度推定に残された大きな課題として，正確な RGB-D 画像の入手不可能性が存在する．Godard ら [13] は自己教師あり学習の概念を用いることでこれを解決し，Unsupervised Monocular Depth Estimation Network:monodepth によって教師なし単眼深度推定手法の確立に貢献した．monodepth は abs rel において 9.7% の推定精度を記録し，教師なし単眼深度推定が注目を集める要因となっている．反面，教師なし学習は学習に使用可能なデータセットの量を増幅させる特徴を持つものの学習に必要なデータの量を削減することはできず，学習コストが高いという欠点を抱えている．以上を踏まえ，本稿では深度勾配を活用した Multi-task learning による教師なし学習モデルを提案する．Multi-task learning により学習に必要なデータセットの量を削減し，定義に主観の存在しない深度勾配を選択することで追加情報による学習への悪影響を排除した上で，教師なし学習の推定精度を向上させることを目的とする．

3 要素技術

3.1 Multi-task learning

Multi-task learning とは機械学習におけるモデルの機械学習手法の一種である．ネットワークを一部共有した多出力ネットワークにおいて複数の関連タスクを同時に解かせることによりタスク間の共通表現を学習させる．これによりモデルの汎化性能や推定精度が高まり，学習や推論の時間ならびにモデルサイズを削減できる．ただし，選択するタスクの組み合わせによっては学習に悪影響を与えるため注意が必要である．本研究では深度と密接な関係を持つ深度勾配を関連タスクとして選択し提案モデルを構築した．

3.2 自己教師あり学習

自己教師あり学習とはモデルに自ら教師データを生成させる機械学習手法の一種である．深度推定においては RGB-D 画像入力として与えずにモデルを訓練する学習のことを指し，単眼深度推定における教師なし学習として広く扱われている．具体的には，中間層の出力に深度を前提とした特殊な推定タスク（他視点画像生成問題やカメラの姿勢推定問題など）をモデルに与えることで，間接的に深度推定を行い，モデルの改善を行

∗ 琉球大学, University of the Ryukyus

う．取得が困難である RGB-D 画像を学習に不要とすることで，学習に利用可能なデータの種類と量を著しく増加させることから，単眼深度推定に有効な手法として注目されている [3]．代表的なモデルとして Godard らの monodepth[13] や Zhou らの SfM Learner[4] が存在する．後者はカメラの姿勢推定問題を単眼深度推定に初めて取り入れた研究の一つであり，これを原型に提案モデルを構築する．

4 先行研究

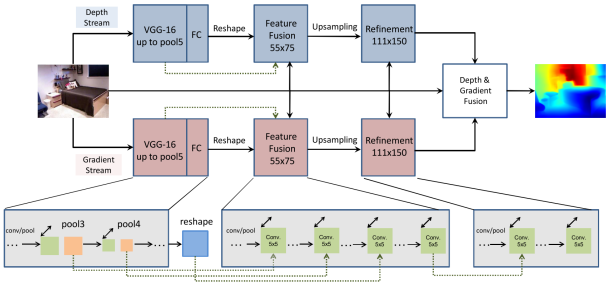

図 1 Two-Streamed network

4.1 Two-Streamed network

Two-Streamed network:Two-Net[6] とは深度勾配の Multiple-input を活用した教師あり単眼深度深度推定モデルである．モデルを図 1 に示す．モデルのネットワークは大きく二つあり，深度を推定する Depth Stream（図 1 上）と深度勾配を推定する Gradient Stream（図 1 下）に分けられる．両者で推定した情報を Multiple-input として CNN ブロックに与えることで最終的な深度を推定する．モデルは深度の推定精度を向上させ深度勾配を追加情報として選択することの優位性を示したが，MSM2[9] のネットワークを二つ内包しているためモデルサイズが大きく，関連性の高いタスクを複数扱っているにも関わらず共通表現を獲得する機構が組み込まれていない．Two-Net を Multi-task learning に対応させることで共通表現を獲得すると同時にモデルサイズを削減し，教師なし学習へ対応させることでより強固なモデルとすることを提案モデルの最終目標とする．

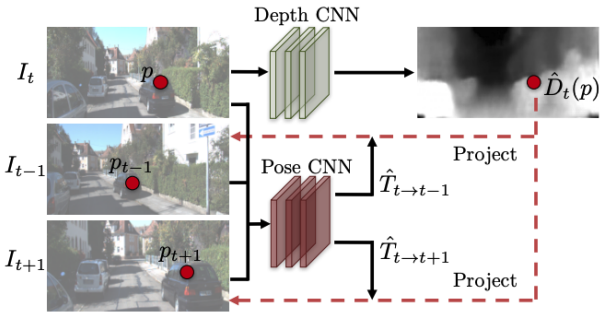

図 2 SfM Learner

4.2 SfM Learner

SfM Learner[4] とは動画像の復元問題を主タスクとして設定した自己教師あり単眼深度推定モデルの一種である．単眼動画像のみから深度推定を学習することを可能

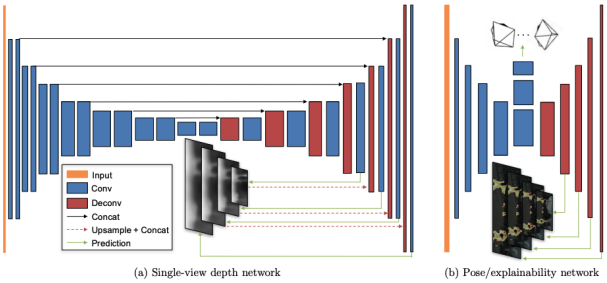

図 3 Depth CNN と Pose CNN[4]

としており，多くの教師なし単眼深度推定モデルの元となっている．モデルの概要を図 2 に示す．カメラの座標を表す自己位置と深度を元にして画面再構成を行い，現在時刻を示す Source image から数秒先の未来を示す Target image を復元することで学習を行う．深度と自己位置の情報は事前に与えられず，画面再構成は両者を用いた数値計算によってのみ求まるため，必然的に，loss の減少に付随して深度が中間層の出力として与えられる．ネットワークの内部構造を図 3 に示す．モデルは深度の出力を想定した Depth CNN と自己位置の推定を想定した Pose CNN の二つのネットワークを持ち，Encoder-decoder に Skip Connection を組み合わせることで効率的に特徴量の抽出と活用を行っている．提案モデルでは Encoder から深度勾配推定部を分岐させることで Multi-task learning を実装し，深度と深度勾配に適した共通表現の獲得を狙う．

5 教師なし学習モデルの提案

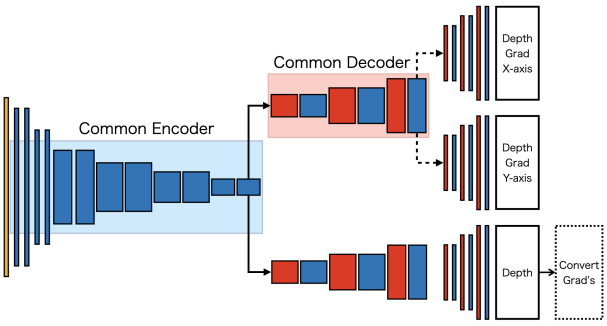

図 4 提案モデル．Skip Connection の描写は省略する．

5.1 提案モデル

提案モデルを図 4 に示す．SfM Learner[4] を雛形として用い，特に言及がない限りはオリジナルモデルに従う．Encoder 以降のネットワークを分岐させることで，x 軸と y 軸に関する深度勾配ならびに深度を同時に推定し学習を行う．従来の Encoder 部である Common Encoder と，深度勾配を推定する直前に配置された Common Decoder を用い，Multi-task learning による共通表現の獲得を狙う．また，推定深度から生成した深度勾配を生成深度勾配とし，推定深度勾配の擬似的な教師データとして活用することで，深度推定部と深度勾配推定部により強い相互作用を与える．Two-Net[6] と比較した場合，Multi-task learning を活用したことによりモデルサイズが削減され，教師なし学習モデルであるため

METHOD	abs diff	abs rel	sq rel	rms	log rms	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Origin	7.9971	0.4747	5.2703	12.5536	0.6165	0.2826	0.5212	0.7243
Our	9.3106	0.6266	8.0841	13.6031	0.7338	0.2249	0.4392	0.6212

表 1　一般的な評価関数による評価

METHOD	recon. loss	smooth. loss	mask loss
Origin	0.1518	0.0433	0.0815
Our	0.1398	0.0206	0.2631

表 2　SfM Learner で提案された loss による評価

に使用できるデータセットの種類が大幅に向上している点で優れている．SfM Learner[4] と比較した場合，提案モデルはより汎用的な特徴量を獲得できることが期待され，学習に使用するデータセットの量を削減できる可能性がある．

6　実験

本実験では提案モデルの学習および精度評価を行い，提案モデルの有用性を検証することを第一目標とする．Zhou らの SfM Learner[4] を精度の比較対照先に設定し，同モデルで提案された loss と一般的な評価関数を用いて評価を行う．なお，特に明記がない場合，使用するモデルの詳細はオリジナルのモデルに準拠する．

6.1　データセット

KITTI データセット [**?**] は KITTI Vision Benchmark Suite が公開する物体検出データセットである．屋外を対象とし，歩行者と車両が含まれた 7400 件以上の学習用画像から構成されている．内容物にはステレオ画像や RGB-D 動画像ならびにラベルと対応した Bounding box などが含まれ，本実験ではこのうちステレオ RGB 動画像を入力として用いている．実験の簡易化のため，Zhou ら [4] が SfM Learner の学習に使用した raw データのうち 2011 年 9 月 26 日に撮影されたデータだけを使用し，augmentation は未使用とした．train データ，validation データ，test データのそれぞれの内訳は以下となる．

- train: 16972 枚
- valid: 7194 枚
- test: 3048 枚

6.2　Loss

Li らの TwoNet[6] の設計を参考に，提案モデルの学習の際には従来の SfM Learner で用いられていた loss に追加して以下の式を与える．ただし，N は総ピクセル数，G_x は推定深度勾配，$\nabla_x D$ は x 軸に関する生成深度勾配を表し，ω は任意に定数を設定できるハイパーパラメータとする．また，$\phi(x)$ は L1 ノルム ($\sqrt{x^2 + 10^{-4}}$) を表している．

$$\omega \sum_{p}^{N} [\phi(\nabla_x D^p - G_x^p) \times \phi(\nabla_y D^p - G_y^p)] \qquad (1)$$

6.3　実験環境

train data の Batch サイズを 4 に設定し，学習開始から 50epoch が経過した時点での精度を評価する．Optimizer に使用した Adam のハイパーパラメーターを表 6.3 に示す．Reconstruction loss，Smoothness loss，Mask loss のハイパーパラメータにはそれぞれ 1,0.1,0.2 を使用した．生成深度勾配の作成には単純な Edge filter ($[-1, 0, 1]$,$[-1, 01]^T$) を用いている．

lr	momentum	beta	weight decay
2×10^{-4}	0.9	0.999	0

表 3　Adam のハイパーパラメーター

6.4　精度評価に使用した関数

6.4.1　一般的な評価関数

実験に使用した評価関数を以下に示す．thresholded accuracy は数値が高いほど優秀だと判断される評価基準であり，推定深度の精密性と外れ値の量を説明する．その他の評価関数は値が低いほど優秀だと判断される評価関数である．

- thresholded accuracy: $\sigma < 1.25, \sigma < 1.25^2, \sigma < 1.25^3$
- mean absolute error: MAE
- squared absolute error: abs. rel
- squared relative error: sqr. rel
- root mean squared error: RMS(lin)
- root mean squared error log_{10}: RMS(log)

6.4.2　SfM Learner で提案された loss

Zhou らは SfM Learner の提案にあたり以下の 3 種類の loss を新たに設計した．recon. loss は画面再構成タスクの精度を表す loss であり，生成画像と Target image との差を absolute error によって表す．smooh. loss は画像中の平坦な箇所に対する制約項である．具体的には近傍ピクセル同士の深度の変化率をペナルティーとして与えている．最後の mask loss は移動物体に対しての masking 領域の広さを表し，Pose Net が自己位置推定の際に出力した不可視領域を元に計算される．mask loss の大小はモデル自身の推定に関する信頼性を間接的に表している．

- reconstruction loss: recon. loss
- smoothness loss: smooth. loss
- mask loss

6.5　結果と考察

表 5.1 と表 5.1 に実験結果をまとめる．thresholded accuracy 以外の関数は値が低いほど優秀であることに注意したい．Origin が SfM Learner を従来手法を用い

て学習した結果，Our が私たちの提案モデルの結果を表す．一般的な評価関数による精度評価では，全ての場合において従来手法に劣る結果となった．対して Zhou ら [4] の提案した loss による精度評価では，画面再構成の精度を表す recon. loss において従来手法を上回る結果となったが，他二つの loss に関しては改善が見られなかった．特に不可視領域の範囲を示す mask loss に関して大きな差が見受けられ，移動物体に対してモデルが過敏に反応していることが読み取れる．また，画面再構成問題の精度を示す recon. loss は推定深度を前提として計算が行われるため，他全ての評価関数に優位性が見られない状況で高精度を記録した事実は直感に反する．これは，recon. loss の計算に使われる自己位置の推定精度が改善されたことで，深度推定に頼らない画面再構成手法をモデルが学んだことによる可能性が高い．深度推定の精度が改善されなかった理由は複数考えられる．ひとつは SfM Learner の loss として設定されている smooth. loss と深度勾配推定が競合した結果，互いの学習に悪影響を与えた可能性である．もう一つは，深度勾配の Multi-task learning が自己位置推定タスクに最適化された結果，深度推定を無視して Pose Net に相互作用を与えてしまった可能性である．教師なし単眼深度推定は原理的に Multi-task learning を必ず内包しているので，教師あり学習と比較して取り扱えるタスクの数に制限があると結論づける．

7 今後の展望

本研究では，深度推定モデルに追加情報を与えることの有効性に着目し，深度勾配推定と Multi-task learning を組み合わせた単眼教師なし学習モデルを提案した．実験の結果，深度勾配推定タスクが自己位置推定タスクに最適化されてしまい，深度推定の精度を改善するには至らなかった．反面，提案モデルを自己位置推定問題へ流用できる可能性が示唆され，空間情報を扱うモデルに深度勾配情報を与えることの有用性が示された．今後の展望として，深度勾配推定問題を smooth. loss の代替として用いた場合の精度を確かめた上で，提案モデルの応用先を検討したい．

謝辞

日頃より熱心なご指導を頂いた御教授ならびに適切な御助言と細かな御配慮を戴いた Lilz 株式会社の Jakub Kolodziejczyk 氏，西銘 大喜氏の両名に深く感謝致します。

参考文献

[1] Wouter Van Gansbeke, Davy Neven, Bert De Brabandere, and Luc Van Gool. Sparse and noisy LiDAR completion with RGB guidance and uncertainty. In *Proceedings of the 16th International Conference on Machine Vision Applications, MVA 2019*, 2019.

[2] Satyarth Praveen. Efficient Depth Estimation Using Sparse Stereo-Vision with Other Perception Techniques, 2019.

[3] Chaoqiang Zhao. Monocular Depth Estimation Based On Deep Learning: An Overview. 2020.

[4] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[5] Omid Hosseini Jafari, Oliver Groth, Alexander Kirillov, Michael Ying Yang, and Carsten Rother. Analyzing modular CNN architectures for joint depth prediction and semantic segmentation. *Proceedings - IEEE International Conference on Robotics and Automation*, No. iv, pp. 4620–4627, 2017.

[6] Jun Li, Reinhard Klein, and Angela Yao. A Two-Streamed Network for Estimating Fine-Scaled Depth Maps from Single RGB Images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[7] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Learning 3-D scene structure from a single still image. In *Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, Vol. 3, pp. 2366–2374. Neural information processing systems foundation, 2014.

[9] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[10] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[11] Lubor Ladický, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[12] Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015.

[13] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Vol. 2017-Janua, pp. 6602–6611, 2017.