H-020

深さ分離畳み込みと多次元協調注意機構を利用した効率的な単眼深度推定
Efficient Monocular Depth Estimation Using Depthwise Separable Convolutions and Multidimensional Cooperative Attention

東内 元気† Genki Higashiuch
嶋田 知泰† Tomoyasu Shimada
孔 祥博‡ Xiangbo Kong
冨山 宏之† Hiroyuki Tomiyama

1. はじめに

近年、人工知能技術の急速な発展により、ロボットやドローン、自動運転車などの製品が人々の生活に広く応用されている。これらの領域において、単眼カメラからの深度マップ生成は、Visual Odometry[1]、三次元再構成[2]、自己位置推定[3]などの基盤情報として不可欠である。深度マップを取得するためには、通常、RGB-D センサ、LiDAR、あるいはステレオカメラが必要とされる。しかしながら、RGB-D センサや LiDAR、ステレオカメラは高コスト・高消費電力であり、ステレオマッチングでは計算負荷やカメラ間の同期ずれが問題であった。このような背景から、ハードウェア要件を抑えつつ低コストで深度マップを取得可能な単眼深度推定が注目を集めている。一方で、深層学習を用いた単眼深度推定は大規模な深度ラベル付きデータを必要とし、その収集には多大なコストと労力を要する。そのため近年では、ステレオペアや単眼動画の再投影誤差を教師信号として活用する自己教師あり手法[4,5]が盛んに研究され、高い性能を達成している。しかしながら、これらのモデルは数千万～数億ものパラメータを有するため、エッジデバイスなど計算資源の限られた環境での実用的な推論は依然困難である。本研究では、Lite-Mono[6]を基盤とした軽量自己教師あり単眼深度推定フレームワークを提案する。提案手法では、デコーダ内の 3×3 畳み込みを深さ分離畳み込み（Depthwise Separable Convolution）[7]に置換して演算コストを大幅削減し、スキップ接続直前に多次元協調注意機構(Multidimensional Collaborative Attention: MCA）[8]を導入して高解像度特徴マップから境界情報を精緻に抽出するとともに、エンコーダ最深部の畳み込みブロックを従来の 10 層から 4 層に削減してモデル全体のパラメータ数および演算量をさらに低減した。KITTI ベンチマーク上の評価では、提案手法が Lite-Mono と比較して学習可能パラメータ数を約 42.1%削減（1.8M）すると同時に、二乗相対誤差を約 4.5%改善した。

2. 提案手法

本節では、はじめに提案手法のアーキテクチャについて説明し、続いて損失関数について述べる。提案手法のアーキテクチャの全体像は図 1 に示す。提案手法は、PoseNet と DepthNet の二つの部分から構成される。PoseNet のエンコーダには事前学習済み ResNet-18 を用い、隣接フレームのカラー画像ペアを入力とする。デコーダは 4 層の畳み込みからなり、隣接画像間の 6-DoF 相対姿勢を推定する。次に、本研究の中核を成す DepthNet の構造について説明する。DepthNet は U-Net に類似したエンコーダ–デコーダアーキテクチャを採用している。エンコーダでは入力画像を段階的にダウンサンプリングしながら多層の特徴抽出を行い、一方デコーダではスキップ接続を介してエンコーダからの特徴を統合しつつ、逐次的に空間解像度をアップサンプリングする役割を担う。

2.1 エンコーダ

提案手法のエンコーダは、4 段階のステージから構成され、入力画像を各ステージで 3×3 畳み込みによって段階的にダウンサンプリングしながら、マルチスケールな特徴抽出と融合を行う。各ステージでは、まず前ステージの出力と平均プーリングした入力特徴を結合し、3×3 畳み込みによって空間解像度を縮小する。その後、CDC モジュール[6]を適用し、多様な受容野を通じて局所的な特徴を連続的に抽出するとともに、LGFI モジュール[6]を挿入して空間軸・チャネル軸の両面から大域的な情報をネットワークに入力する。Stage 2·3 では、前ステージからのダウンサンプリング出力を残差接続的に連結し、ステージ間の特徴相関を強化した上で次段階へ引き継ぐことで、最終的により高次な抽象特徴を獲得する。提案手法では、Stage 4 の畳み込みブロック数を 10 層から 4 層に削減し、モデル全体のパラメータ数を削減しつつ、従来同等の推論精度を維持している。

2.2 デコーダ

提案手法のデコーダは、3 段階のアップサンプリングと畳み込みブロックペアから構成され、各ステージでマルチスケールの深度マップを生成する。まず、前段から受け取った特徴マップを標準的な ConvBlock（3×3 畳み込み＋バッチ正規化＋ReLU）で処理し、バイリニア補間により解像度を 2 倍に拡大する。次に、対応ステージのエンコーダ特徴をスキップ接続で結合した後、Depthwise Separable Convolution を適用する。Depthwise Separable Convolution は、3×3 の Depthwise Convolution でチャネル毎の空間情報を効率的に抽出し、1×1 の Pointwise Convolution でチャネル間を統合する構造で、従来手法である Lite-Mono の 2 連続の 3×3 畳み込みに比べて演算量とパラメータ数を大幅に削減しながら、高解像度エッジやテクスチャを保持することが出来る。最終的に各スケールごとに 1×1 畳み込みで深度マップを出力し、Sigmoid によって正規化する。これにより、エッジ検出精度と計算効率の両立を実現している。

2.3 MCA

MCA は、空間軸（幅・高さ）およびチャネル軸という三つの独立した次元に対して注意重みを計算し，それぞれの注意マップを統合することで，局所的な輪郭情報とチャネル間の意味的な特徴を同時に強調する手法である。まず

† 立命館大学 Ritsumeikan University
‡ 富山県立大学 Toyama Prefectural University

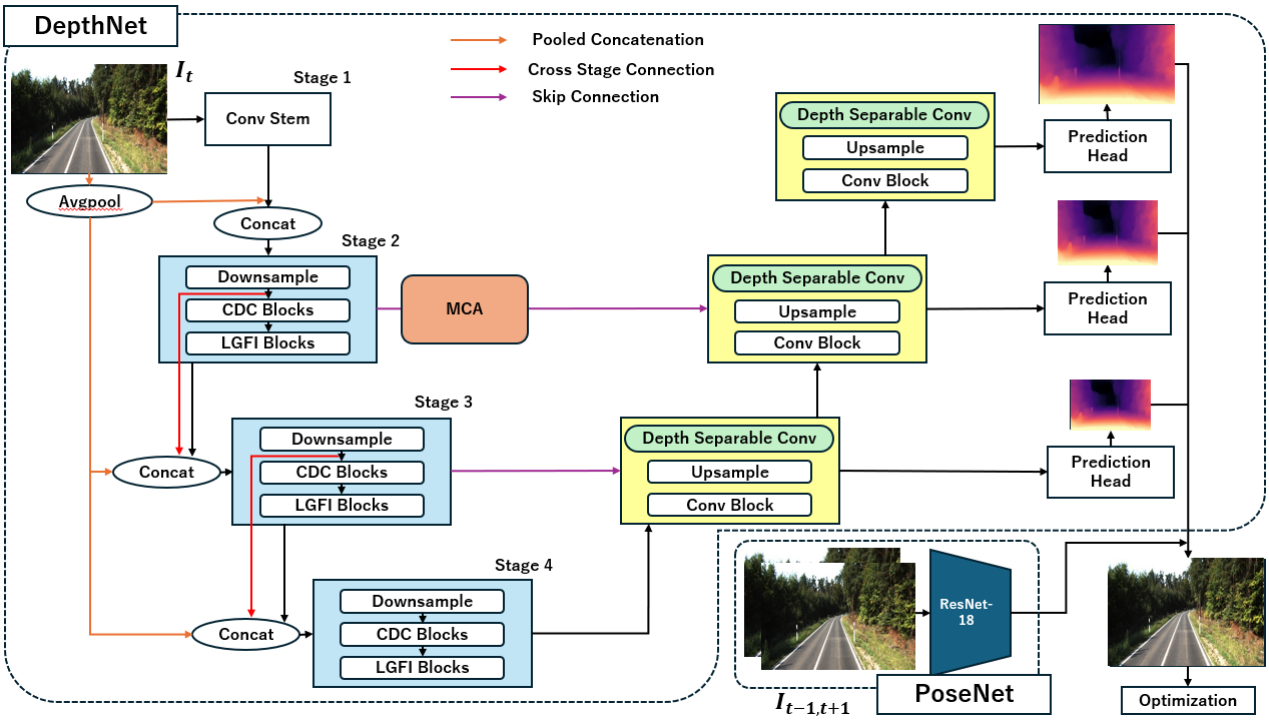

図 1: 提案手法のアーキテクチャの全体像

表 1: 提案手法を Lite-Mono と比較した結果

	誤差（↓）			精度（↑）			（↓）	
手法	Abs Rel	Sq Rel	RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	FLOPs (G)	Params (M)
Lite-Mono	0.109	0.872	4.712	0.885	0.961	0.982	5.032	3.1
提案手法	0.11	0.833	4.644	0.883	0.962	0.982	4.008	1.8

幅方向・高さ方向それぞれの自己相関行列を構築して空間的特徴を抽出し，次にチャネル軸全体にグローバルプーリングを適用してチャネルごとの重要度を推定する。最後に，これら三つの注意マップを融合し，1×1 畳み込みと非線形活性化を経て元のチャネル次元に復元することで，高解像度な特徴表現を実現している。提案手法では、エンコーダの Stage 2 の出力をデコーダに渡すスキップ接続直前に MCA を挿入し、高解像度特徴マップが保持する細かな輪郭情報と意味的特徴を同時に強化する。エンコーダの Stage 2 の出力はダウンサンプリングが少なく、物体輪郭や微細構造を豊富に含むため、MCA の多次元的な注意重み付けが追加パラメータをほとんど増やさずに高精度な深度推定へと寄与する。

2.4 損失関数

提案手法の最終的な損失関数Lは、Lite-Mono と同様にマルチスケール出力$s \in \left\{1, \frac{1}{2}, \frac{1}{4}\right\}$ごとに、画像再構成損失$L_r^s$とエッジ認識スムースネス損失$L_{smooth}^s$を計算し、経験的に設定した重み係数$\lambda = 10^{-3}$を用いて平均化することで定義される。$L_r^s$はスケール$s$におけるターゲット画像と再投影合成画像間の光度再投影損失を示し、L_{smooth}^sは同スケールでの逆深度マップの勾配に対しエッジ保存性を考慮した平滑化を行う損失である。

$$L = \frac{1}{3}\sum_{s\in\left\{1,\frac{1}{2},\frac{1}{4}\right\}}\left(L_r^s + \lambda L_{smooth}^s\right) \quad (1)$$

3. 実験

本節では、提案手法の有効性を検証し、既存手法との比較によりその優位性を示す。

3.1 実験環境

3.1.1 データセット

本研究では、KITTI データセット[9]を用いる。KITTI は 61 シーンのステレオ道路走行データを収録し、自動運転やロボティクス研究向けにカメラ、LiDAR、GPS、IMU など複数のセンサで取得された実車映像・計測情報を含む。学習および評価には、Eigen らによる標準スプリット[10]を採用し、学習用に 39,180 組の単眼トリプレット、検証用に 4,424 フレーム、テスト用に 697 フレームを用いる。自己教師あり学習では、既知のカメラ内部行列を利用し、データセット全体の焦点距離の平均値をすべての学習画像に共通の内部パラメータとして設定する。また評価時には、慣例に従い予測深度を 0〜80 m の範囲にクリップする。

3.1.2 性能評価

KITTI データセットにおける性能評価には、深度推定分野で広く採用されている 6 つの指標を用いる。具体的には、絶対相対誤差（Absolute Relative Error: Abs Rel）、二乗相対誤差（Squared Relative Error: Sq Rel）、および二乗平均平方根誤差（Root Mean Squared Error: RMSE）を計測し、加え

て閾値精度として$\delta < 1.25$、$\delta < 1.25^2$、$\delta < 1.25^3$の 3 つの指標を算出する。それぞれ次式で定義する。

$$Abs\ Rel = \frac{1}{|N|}\sum_{i \in N}\frac{[d_i - d_i^*]}{d_i^*} \quad (2)$$

$$Sq\ Rel = \frac{1}{|N|}\sum_{i \in N}\frac{\|d_i - d_i^*\|_2}{d_i^*} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{|N|}\sum_{i \in N}\frac{\|d_i - d_i^*\|_2}{d_i^*}} \quad (4)$$

$$Acciracies = max\left(\frac{d_i}{k_i^*}, \frac{d_i^*}{d_i}\right) = \delta < threshold \quad (5)$$

ここで、Nは深度の正解データを持つ画素の総数、d_iは画素iにおける推定深度、d_i^*は画素iにおける正解深度を表す。また、モデルの軽量性と高速性を評価するため、学習可能パラメータ数および FLOPs も併せて計測している。

3.2 実験結果

提案手法を Lite-Mono と比較した結果を表 1 に示す。提案モデルは Lite-Mono に対し、FLOPs を約 20.3%削減し、パラメータ数を約 42.1%（1.8M）削減しながら、全 6 指標において平均 0.85%の性能向上を達成している。中でも Sq Rel は 4.47%の改善、RMSE は 1.44%の改善を示した。提案手法では、デコーダ内の 3×3 畳み込みを Depthwise Separable Convolution に置き換えたことで、ほとんど性能を落とさず軽量化できることが確認された。またエンコーダの出力をデコーダに渡すスキップ接続直前に MCA を挿入することで、軽量化後も物体輪郭や意味情報を強化し、推論精度と処理効率を同時に高めている。さらに、エンコーダの Stage4 の畳み込みブロック数を削減することで性能を維持しつつ、モデル規模と計算コストの大幅な低減を達成できた。

4. おわりに

本研究では、軽量かつ高精度な自己教師あり単眼深度推定を実現するアーキテクチャを提案した。具体的には、デコーダ内の従来 3×3 畳み込みを Depthwise Separable Convolution へ置き換え、スキップ接続直前に MCA を導入し、さらにエンコーダ最深部の畳み込みブロック数を削減するという 3 つの改良を加えた。KITTI ベンチマーク上の評価結果では、Lite-Mono と比較して約 42%のパラメータ削減と約 1.4%の RMSE 改善を同時に達成し、リソース制約下のエッジ環境においても高精度な深度推定が可能であることを示した。今後の課題としては、さらなる推論速度の最適化を通じたリアルタイム性能の向上が挙げられる。

謝辞

本研究の一部は公益財団法人スズキ財団の科学技術研究助成、および NEDO の委託(JPNP22006)による。

参考文献

[1] D. Nistér, O. Naroditsky, and J. Bergen, “Visual Odometry,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.652–659, 2004.
[2] Z. Ma and S. Liu, “A Review of 3D Reconstruction Techniques in Civil Engineering and Their Applications,” Advanced Engineering Informatics, vol. 37, pp.163–174, 2018.
[3] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, “MonoSLAM: Real-Time Single Camera SLAM,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1052–1067, 2007.
[4] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, “Unsupervised Learning of Depth and Ego-Motion from Video,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1851-1858, 2017.
[5] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, “Digging into Self-Supervised Monocular Depth Estimation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, pp. 3828-3838, 2019.
[6] N. Zhang, F. C. Nex, G. Vosselman, and N. Kerle, “Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 18537-18546, 2023.
[7] F. Chollet, “Xception: Deep learning with Depthwise Separable Convolutions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251-1258, 2017.
[8] Y. Yu, Y. Zhang, Z. Cheng, Z. Song, and C. Tang, “MCA: Multidimensional Collaborative Attention in Deep Convolutional Neural Networks for Image Recognition,” Engineering Applications of Artificial Intelligence, vol. 126, Art. 107079, 2023.
[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision Meets Robotics: The KITTI dataset,” The International Journal of Robotics Research, vol. 32, no. 11, pp. 1231-1237, 2013.
[10] D. Eigen and R. 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, Santiago, pp. 2640–2648, 2015.