2U-04

Vision Transformerの単眼深度推定への応用

石川泰暉　長名優子

東京工科大学 コンピュータサイエンス学部

1　はじめに

近年、Transformer[1] アーキテクチャの利用が急速に拡大している。Transformer はその卓越した表現力と柔軟性により、自然言語処理だけでなく、画像処理分野でも驚くべき性能を発揮している。例えば、2020 年に発表された Vision Transformer[2] は、Transformer の Encoder 部分を画像分類タスクに応用したものである。このことはコンピュータビジョンの分野においてもブレイクスルーとなった。近年、Vision Transformer をベースとしたモデルを用いた手法が画像分類だけでなく、セグメンテーション等の多くのタスクにおいても認識精度の最高スコアを更新しており、Transformer アーキテクチャの汎用性が広く認識されている。本研究において扱う単眼深度推定は、単一のカメラ画像から 3D 空間の距離情報を推定する重要なタスクであり、自動運転技術の進展や AR(拡張現実) アプリケーションの進化において非常に重要な役割を果たしている。従来の手法では、複数のカメラを利用するか、専用のセンサーを使用する必要があったが、機械学習の登場により、単一のカメラ画像のみから情報を取得することが可能になった。単眼深度推定においても ViT をベースとしたモデルは複数発表されており、その多くが高い予測スコアを記録している。しかし、高精度を達成するためのモデルの複雑化という問題は避けられず、計算量の増加を招いている。

本研究では、これらの課題に対処するため、Vision Transformer をバックエンドに用いた可能な限りシンプルなモデルで単眼深度推定を行うことを試みる。

2　単眼深度推定

単眼深度推定は、1 つのカメラ (単眼カメラ) の画像から物体の深さ情報を推定する技術であり、2 次元の画像から 3 次元の深さ情報を推定することを目的とする。単眼深度推定のネットワークは、一般的に 3 チャンネルのカラー画像または 1 チャンネルのグレースケール画像を入力とし、1 チャンネルの深度マップを出力する。入力画像は RGB 情報を持ち、推定された深度マップは画像内の各ピクセルに対応する 3D 空間上の距離を示す。単眼深度推定の特徴として、推論時にリアルタイム性が求められるケースが多い点が挙げられる。たとえば、自動運転では迅速な判断が必要であり、スマートフォンのカメラアプリケーションでもリアルタイムな映像処理がユーザエクスペリエンスを向上させる。そのため、高速な推論が可能なネットワークの開発が重要な課題となっている。身近な利用分野としては、自動運転やスマートフォンのポートレート撮影などがある。自動運転技術では、単眼カメラを装備した車両が周囲の道路や障害物の距離を正確に把握することで、より安全な運転が可能となる。また、スマートフォンのポートレートモードでは、被写体と背景の距離を推定して、美しいボケ効果を実現することができる。このように、単眼深度推定は、現代のコンピュータビジョン技術の中でも特に注目すべき重要な研究分野であり、その応用範囲や精度の向上によって、さまざまな領域に革新をもたらすことが期待されている。

3　Vision Transformerの単眼深度推定への応用

ここでは、提案する Vision Transformer を用いた単眼深度推定手法について説明する。

提案手法では、ViT Encoder と CNN[3] Decoder の 2 つからなるエンコーダ・デコーダモデルを採用しており、ViT エンコーダでエンコードした情報を CNN Decoder で元の大きさまでアップスケーリングしながら復元することにより、最終的な深度マップ出力を生成する。

ViT Encoder は ViT[2] モデルをベースに設計している。ただし、本研究では分類タスクではなく回帰タスクを扱うため、class token や MLP Head がない点で元のモデルと異なっている。また、計算量を削減するために原論文において提案された ViT-Base よりもモデルの層数や次元数を小さくしている。

CNN Decoder は GAN[4] の Generator を参考に設

Application of Vision Transformer to Monocular Depth Estimation

Taiki Ishikawa and Osana Yuko(Tokyo University of Technology, osana@stf.teu.ac.jp)

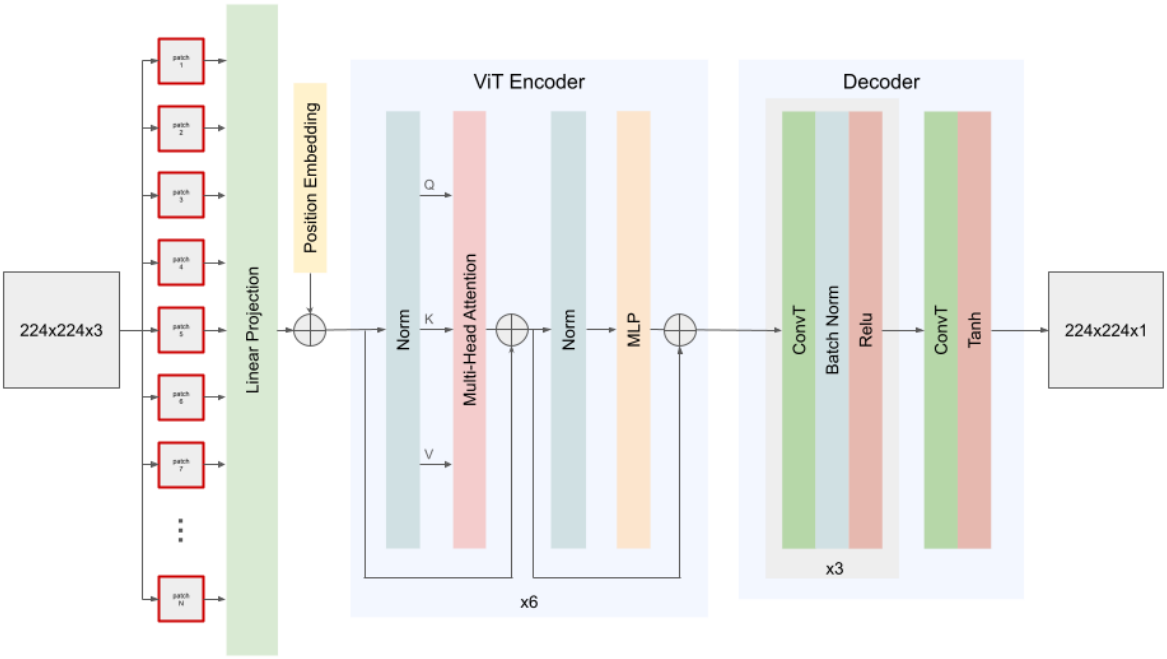

図 1: 提案手法で用いるネットワークの構造

計しており、転置畳み込み、バッチ正規化、Relu アクティベーションの 3 層を 3 回、転置畳み込みと Tanh アクティベーションの 2 層を 1 回繰り返す構造になっている。ここで、アップスケーリング層の代わりに転置畳み込みを使用したのは、転置畳み込みの方がアップスケーリングと比較したときの計算量が小さいためである。

提案手法で用いるネットワークの構造を図 2 に示す。

4 計算機実験

計算機実験を行い、提案手法のようなシンプルな構造でもある程度単眼深度推定が行えることを確認した。図 2 にその結果の一部を示す。

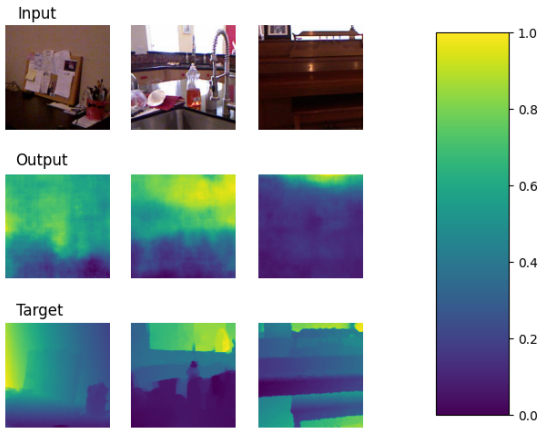

図 2: 単眼深度推定結果

参考文献

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin : “Attention is all you need,” https://arxiv.org/pdf/1706.03762.pdf, 2017 (2024/01/10 参照).

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby : “An Image is worth 16x16 words: Transformers for image recognition at scale,” Proceedings of International Conference on Learning Representations, 2021.

[3] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner : “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, Vol.86, No.11, pp.2278–2324, 1998.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio : “Generative adversarial nets,” Neural Information Processing Systems, pp2672–2680, 2014.