H-005

Deform-Conv を適用した HRNet による可変受容野 Semantic Segmentation
Variable receptive field Semantic Segmentation by HRNet using Deform-Conv

安藤 大貴 † 荒井 秀一 †
Daiki Ando Shuichi Arai

1. はじめに

Semantic Segmentation はピクセル単位でクラス分類を行う技術である．この技術の登場により，それまでの物体検出では識別できなかった物体の詳細な形状を識別可能になった．物体の形状は，物を掴むなど，物に実際に触れる際に必要な情報である．ゆえに，Semantic Segmentation は画像認識分野で研究が盛んであり，ロボットビジョンへの応用も期待されている．

Semantic Segmentation の手法として，Bottleneck 構造の CNN(Convolutional Neural Network) が多く用いられた．これらの手法では，ネットワークの浅い層では物体の形状などの細かな特徴を抽出し，深い層ではより抽象的な特徴を抽出する．この抽象化のために，多くの場合ダウンサンプリングにより画像のサイズを落とすが，この処理により細かな特徴が失われてしまう問題があった．近年では，この問題を解決するために，高解像度から低解像度の複数解像度を並列に処理する “HRNet(High-Resolution Network)” [1] が登場し，識別精度は飛躍的に向上した．しかし，依然として画像の占有面積が小さい物体は，他クラスと比較して低い問題があった．

2. 提案

本稿では，画像の占有面積が小さい物体の識別精度が他クラスと比較して低い原因は Convolution にあると考えた．実際の画像中には様々な形状，大きさの物体が存在するが，それらに関わらず受容野が固定形状，固定サイズの Convolution では物体細部の特徴を表現し切れないと考える．そこで受容野の形状，サイズが入力に応じて動的に変化する “Deformable Convolution” [2] を HRNet に導入することを提案する．

2.1.HRNet(High-Resolution Network)

HRNet は異なる解像度を処理するネットワークを組み合わせた 4 つのステージで構成されている (図 1)．ステージ 1 は入力と同等の解像度で処理するネットワークのみで構成される．続くステージ 2 ではステージ 1 のネットワークに，入力の 1/2 倍スケールの解像度を処理するネットワークを追加する．さらに続くステージ 3 では 1/4 倍スケールの解像度を処理するネットワークを追加し，最後のステージ 4 では 1/8 倍スケールの解像度を処理するネットワークを追加する．

2.2.Deformable Convolution (Deform-Conv)

Deformable Convolution は Convolution のサンプリング位置からの変位である “offset” を計算，学習することでフィルタカーネルのサンプリング位置を動的に変化させる．そのため，学習が進むにつれて画像中の物体

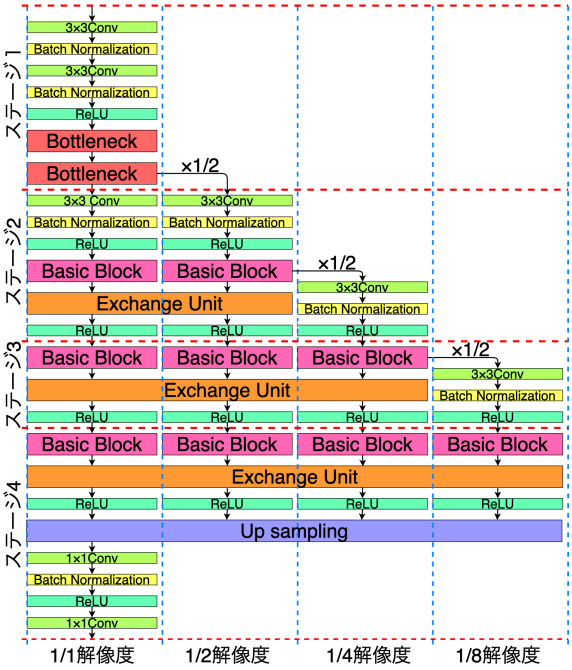

図 1: HRNet

のサイズ，形状に合わせて受容野のサイズ，形状を変えることが可能である．Convolution のサンプリング位置 p_n の集合を R，中心位置を p_0，重みを w，入力特徴マップを x，出力特徴マップを y としたとき，Convolution は式 (1) と表せる．

$$y(p_0) = \sum_{p_n \in R} w(p_n) x(p_0 + p_n) \tag{1}$$

Deformable Convolution では offsetΔp が加わり，式 (2) と表せる．

$$y(p_0) = \sum_{p_n \in R} w(p_n) x(p_0 + p_n + \Delta p) \tag{2}$$

offset は図 2 に示すように，入力特徴マップに追加の Convolution により求め，学習する．

2.3. 提案手法 (HRNet + Deform-Conv)

HRNet には “Basic Block” (図 3 左) という特徴抽出を担う重要な箇所がある．この Basic Block の Convolution(図 3 中の 3×3Conv) の一部を Deformable Convolution(図 3 中の 3×3Deform-Conv) に置き換え，これを “Deformable Basic Block” (図 3 右) と呼ぶことにする．また，Deformable Basic Block は HRNet のステージ 4, 1/8 解像度部分に適用した．

†東京都市大学 総合理工学研究科

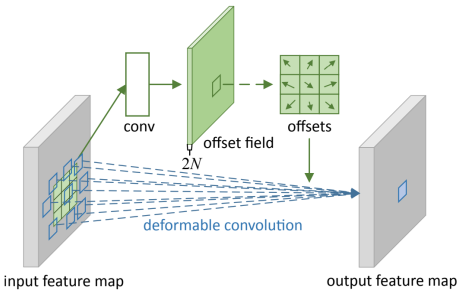

図 2: Deformable Convolution の流れ [2]

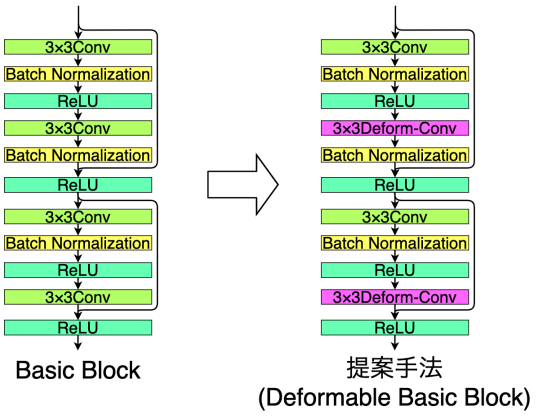

図 3: Basic Block と Deformable Basic Block の構造

3. 実験及び実験結果

実験には, Cityscapes[3] データセットを使用した. 学習用に 2975 枚, 検証用に 500 枚, テスト用に 1525 枚用意されており, 評価対象のクラスは 19 クラスである. 本稿では学習用 2975 枚で学習し, 検証用 500 枚で推論を行った. 提案手法の有効性を示すために, Semantic Segmentation で主に用いられている評価指標 MIoU(Mean Intersection over Union) で評価した. また, HRNet には様々な規模のモデルが提案されており, 本稿では異なる 3 つの規模の HRNet に Deformable Basic Block を適用した. MIoU による評価結果を表 1 に示す.

表 1: 従来手法と提案手法の比較

model	MIoU
HRNet W18-samll v1	70.3
提案手法 (HRNet W18-samll v1)	74.2 (+3.9)
HRNet W18-small v2	76.2
提案手法 (HRNet W18-small v2)	77.6 (+1.4)
HRNet W48	80.5
提案手法 (HRNet W48)	81.0 (+0.5)

表 1 より, 提案手法は従来手法と比較して MIoU は向上している.

次に, 特に変化が大きくみられた, HRNet W18-small v1 と提案手法 (HRNet W18-small v1) のセグメンテーション結果を図 4 示す.

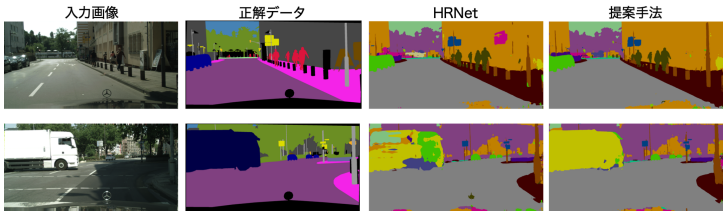

図 4: セグメンテーション結果の比較

図 4 より, 従来手法である HRNet では, pole が途切れてしまっているのに対し, 提案手法では途切れずに識別できている. また truck のような大きな物体の識別結果も大幅に改善している. これは Deformable Convolution により物体のサイズ, 形状を捉えられるようになったためだと推測する.

4. おわりに

本稿では, 複数解像度を並列に処理する HRNet においても画像の占有面積が小さい物体の識別精度が他クラスと比較して低く, その原因が受容野が固定形状, 固定サイズの Convolution にあると考えた. そこで受容野のサイズ, 形状が可変な Deformable Convolution を HRNet に導入する手法を提案した. その結果, MIoU の評価指標による従来手法との比較から, 提案手法の有効性を確認した.

参考文献

[1] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019.

[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.