Realtime Multi-Person Pose Estimation with RCNN and Depthwise Separable Convolution

Van-Thanh Hoang
School of Electrical Engineering
University of Ulsan
Ulsan, Korea
Email: thanhhv@islab.ulsan.ac.kr
ORCID: 0000-0003-3478-9954

Van-Dung Hoang
Department of Engineering-Technology
Quang Binh University
Quang Binh, Vietnam
Email: dunghv@qbu.edu.vn
ORCID: 0000-0001-7554-1707

Kang-Hyun Jo
School of Electrical Engineering
University of Ulsan
Ulsan, Korea
Email: acejo@ulsan.ac.kr
ORCID: 0000-0001-8317-6092

***Abstract*—Human pose estimation is a fundamental research topic in computer vision. This topic has been largely improved recently thanks to the development of the convolution neural network. This paper introduces an efficient human pose estimator based on Mask RCNN. It uses MobileNetV3 as backbone and replaces the vanilla convolutions with the expanded depthwise separable convolutions to reduce the model size, FLOPs and inference time. The model can run in realtime speed at 25 FPS with acceptable scores.**

***Index Terms*—human pose, depthwise separable convolution, RCNN, realtime**

I. Introduction

The multi-person pose estimation problem is to recognize and locate the position of key-points of all people in the image. This task can be applied in many applications like 3D pose estimation [8], [9], human-system interaction, human action recognition/prediction, and video surveillance system.

Recently, the problem of estimation pose of all persons in the image has been greatly improved by the development of the convolution neural networks (CNN) [14]. For example, the Convolution Pose Machine proposed by Cao et al. [2] tried to locate the position of the key-point joints and the connection between them which called part affinity fields (PAFs) in the image. And then ensembles these joints into the full pose of every people inside the image. The Stacked Hourglass Network [17] uses a human detector to have the bounding boxes of every people in the input image. Then, for each person, it generates the score-maps for every key point by using a stack of eight Hourglass modules. Mask-RCNN [6] predicted the bounding box of all persons first, then warps the feature maps based on these boxes to obtain the key-points for the person inside.

This paper modifies the Mask RCNN to have a smaller model, which can run in realtime speed. It uses MobileNetV3 [10] as backbone instead of Residual Network [7], and replaces the vanilla convolutions with the depthwise separable convolutions to reduce the model size, FLOPs and inference time. Additionally, this paper introduces the expanded depthwise convolution to improve the performance of depthwise convolution with small increment cost in model size, FLOPs and inference time. The experiments show that the new model can run in realtime speed at 25 FPS with acceptable mAP scores.

II. Related work

Human pose estimation is a very active research field in computer vision for decades. Classical approaches [4], [16], [20], [25], [28] tackling this problem as a tree-structured or graphical model problem and predict keypoint locations based on hand-crafted features.

Recent works [1], [13], [17] mostly base on the convolution neural network (CNN) [14], which sharply improve the performance of not only human pose estimation but also other tasks in computer vision research. This paper mainly focuses on methods based on CNN. This topic can be categorized as single-person pose estimation which predicts the location of joints of a human with given bounding box, and multi-person pose estimation that requires further recognition of the poses of all humans in the image. Because the multi-person pose estimation task has a high demand for real-life applications, it is gaining increasing popularity recently. The approach of the multi-person case can be divided into two categories: the bottom-up approach and the top-down approach.

Bottom-up approach. Methods that use bottom-up approach will directly predict all key-points at first and then ensemble them into full poses of all persons in the image. Cao et al. [2] encoded the relationship between key-points by using part affinity fields (PAFs) and assemble detected key-points into full poses of different people. This method is integrated into the OpenPose library. DeepCut [21] translated the problem of separating key-points of different people in an image as an Integer Linear Program (ILP) problem. They firstly partitioned the part candidates into person clusters, then combined these clusters with labeled body parts. DeeperCut [13] is an improved version of DeepCut by employing image-conditioned pairwise terms and a deeper ResNet [7] as the backbone network to get better performance.

Top-Down approach. Methods that use top-down approach will interpret the process of estimating pose of all people as a two-step pipeline. That is, firstly detect the bounding box for persons in the image, and then solve the single person pose estimation problem in the cropped patches that based on

TABLE I
DEPTHWISE SEPARABLE CONVOLTUION TRANSFORMING FROM d_i TO d_j CHANNELS, WITH STRIDE s.

Input	Operator	Output
$h \times w \times d_i$	DWConv	$\frac{h}{s} \times \frac{w}{s} \times d_i$
$\frac{h}{s} \times \frac{w}{s} \times d_i$	1×1 conv2d	$\frac{h}{s} \times \frac{w}{s} \times d_j$

TABLE II
EXPANDED DEPTHWISE SEPARABLE CONVOLTUION TRANSFORMING FROM d_i TO d_j CHANNELS, WITH EXPAND FACTOR e AND STRIDE s.

Input	Operator	Output
$h \times w \times d_i$	DWConv	$\frac{h}{s} \times \frac{w}{s} \times ed_i$
$\frac{h}{s} \times \frac{w}{s} \times ed_i$	1×1 conv2d	$\frac{h}{s} \times \frac{w}{s} \times d_j$

the corresponding bounding boxes. Mask-RCNN [6] predicts human bounding boxes first, and then predict human key-points based on the cropped feature map of the corresponding bounding box. Papandreou et al. [18] predict both heatmaps and offsets of the points on the heatmaps to the real positions, and then uses both of them to obtain the final predicted location of key-points.

This paper modifies the Mask RCNN to have smaller model, which can run in realtime speed. It uses MobileNetV3 [10] as backbone instead of Residual Network [7]. Moreover, the vanilla convolutions are replaced by the depthwise separable convolutions to reduce the model size, FLOPs and inference time. Additionally, this paper introduces the expanded depthwise convolution to improve the performance of depthwise convolution with small increment cost in model size, FLOPs and inference time. The experiments show that the new model can run in realtime speed at 25 FPS with acceptable scores.

III. OUR APPROACH

A. Expanded Depthwise Separable Convolution

1) Depthwise Separable Convolution: Nowadays, there are many efficient neural network architectures [3], [11], [24], [29] use Depthwise Separable Convolutions (DWConv) as the key building block. The basic idea of DWConvolution is to replace a standard convolutional layer with two separate layers. The first layer uses a depthwise convolution operator. It applies a single convolutional filter per input channel to capture the spatial information in each channel. Then the second layer employs a pointwise convolution, which is a 1×1 convolution, to capture the cross-channel information.

Suppose the input tensor L_i has size $h \times w \times d_i$, the output tensor L_j has size $h \times w \times d_j$. So, the standard Convolution needs to apply a convolutional kernel $K \in \mathcal{R}^{k \times k \times d_i \times d_j}$, where k is the size of kernel. Therefore, it has the computation cost of $h \cdot w \cdot d_i \cdot d_j \cdot k \cdot k$.

In case of DWConv, the depthwise convolution layer costs $h \cdot w \cdot d_i \cdot k \cdot k$ and the 1×1 pointwise convolution costs $h \cdot w \cdot d_i \cdot d_j$. Hence, the total computational cost of DWConv is $h \cdot w \cdot d_i \cdot (k^2 + d_j)$. Effectively, the computational cost of DWConv is smaller than the standard Convolution by a factor of $\frac{k^2 \cdot d_j}{(k^2 + d_j)}$. The transformation of DWConv is shown in Table I.

2) Expanded Depthwise Separable Convolution: This paper introduces an upgraded of DWConv called Expanded Depthwise Convolution (EDWConv). Instead of applying a single convolution filter per input channel, it applies e convolution filters per input channel to capture multiple spatial information in each channel. Similar to DWConv, it is followed by a pointwise convolution to capture the cross-channel information.

About the computation cost of EDWConv, the expanded depthwise convolution costs $h \cdot w \cdot e \cdot d_i \cdot k \cdot k$ and the 1×1 pointwise convolution costs $h \cdot w \cdot e \cdot d_i \cdot d_j$. Therefore, the total computational cost of EDWConv is $h \cdot w \cdot e \cdot d_i \cdot (k^2 + d_j)$. Effectively, the computational cost of DWConvolution is smaller than the standard Convolution by a factor of $\frac{k^2 \cdot d_j}{e \cdot (k^2 + d_j)}$. Table II illustrates the transformation of EDWConv.

B. Network Architecture

As shown in Fig. 1, the proposed model consists of four parts: the MobileNetV3 [10] as the feature extractor, the RPN to extract the proposals bounding box, the Detection Network to predict the final bounding boxes of all human inside the input, and the Key-points Estimation Network to generate the key point positions of the person inside.

MobileNetV3 [10] is an improved version of MobileNet [11] and MobileNetV2 [24]. It uses inverted residual blocks combined with incorporates squeeze-and-excitation blocks [12] as part of the search space. Its architecture is found out by adopted two AutoML techniques: MnasNet [26] and NetAdapt [27]. MobileNetV3 first searches for a coarse architecture using MnasNet, which uses reinforcement learning to select the optimal configuration from a discrete set of choices. After that, the model fine-tunes the architecture using NetAdapt, a complementary technique that trims under-utilized activation channels in small decrements.

Faster RCNN [22] has two parts: the RPN subnet for generating the box proposals, and then, for each proposal, it uses using a region of interest pooling (ROIAlign) operator to extract a fixed-size feature vector from the feature map provided by the network backbone (MobileNetV3 in this paper). Each feature vector is fed into a sequence of fully connected layers (Detection Network) that finally refine the proposals to get better bounding box and classification results.

Key-points Estimation Network use the bounding boxes generated from Faster RCNN to adopt the ROIAlign operator to extract a fixed-size feature vector from the feature map provided by the network backbone. Each feature vector is fed into a sequence of convolution layers (which can be vanilla Convolution, DWConv, EDWConv).

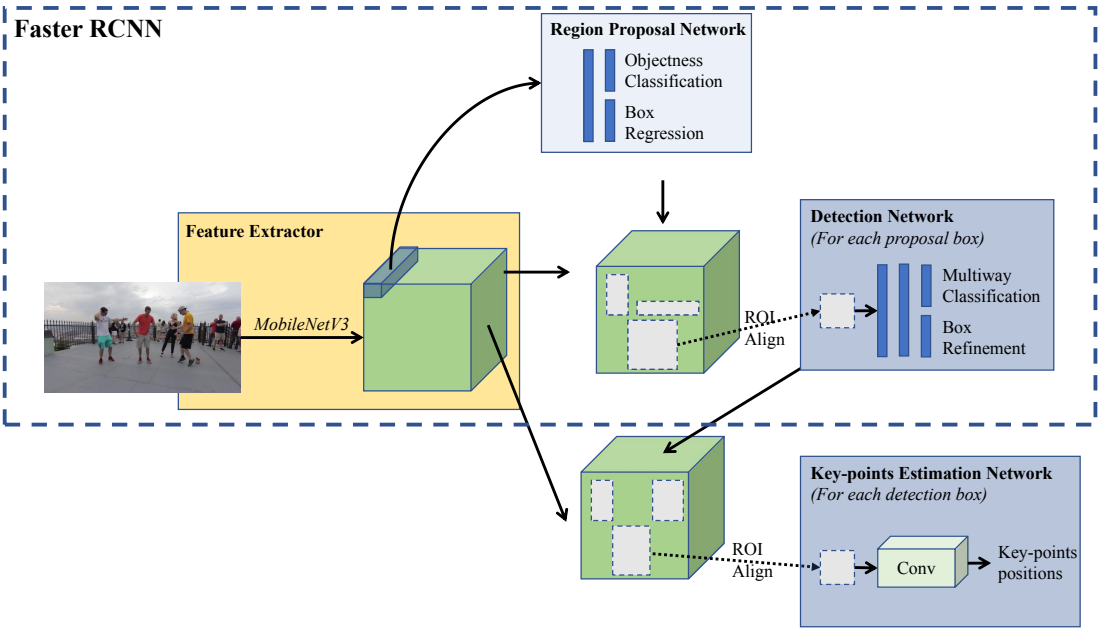

Fig. 1. Overview architecture of the proposed model. It has four parts: the MobileNetV3 [10] as the backbone, the RPN to extract the proposals bounding box, the Detection Network to predict the bounding boxes of all human inside the input, and the Key-points Estimation Network to generate the key point positions.

IV. Experiments

A. Implementation Details

The proposed model is trained on the Coco dataset [15]. This dataset is composed of 118k images (train2017) for training and 5k images (val2017) for validation. This dataset can be used for many tasks, e.g. human pose estimation, object detection. For human pose estimation task, there are 17 joints are annotated. They are: nose, right/left shoulder, right/left elbow, right/left wrist, right/left hip, right/left knee, right/left ankle, right/left eye, and right/left ear.

For the MobileNetV3 backbone, a pre-trained model that already trained on ImageNet [23] dataset is used. The remained parameters of RPN, Detection Network, Key-points Estimation Network are initialized by Xaviers initializer [5].

All the input images are resized to 320×320 pixels. The network is trained using Pytorch [19] and for optimization, the SGD optimizer with a learning rate of 2e-2 is used. The models are trained on a server with a Core i7-8700K 3.70GHz CPU, 32-GB RAM, and 2 NVIDIA Titan RTX GPU devices for 90,000 iterations. The learning rate is dropped once by a factor of 10 at epochs of 60,000 and 80,000.

B. Experiment Results

Table III shows the AP, AR scores and speed of proposed models with several kinds of convolution: vanilla Convolution, Depthwise Separable Convolution, Expanded Depthwise Separable Convolution for the Key-points Estimation Network on Coco dataset.

As can be seen, after replacing vanilla convolutions with depthwise separable convolution, although the AP and AR scores are downgraded (from 46.1 and 54.8 to 37.8 and 46.4, respectively), the speed of network is double. It can achieve 25.2 FPS with acceptable scores, good enough for the realtime application. When adopting the expanded depthwise separable convolution to the Key-points Estimation Network, the model is a little bit slower (25.2 FPS down to 25 FPS) but can have approximately 1% score increasing in both AP and AR.

Some visual examples of multi-person poses predicted by the proposed model are shown in Fig. 2. As you can see, the proposed model can generate good poses for all people inside the input image.

V. Conclusion

This paper modifies the Mask RCNN to have smaller model. It uses MobileNetV3 as backbone instead of Residual Network and replaces the vanilla convolutions with the expanded depthwise separable convolutions to reduce the model size, FLOPs and inference time. The experiments show that the new model can run in realtime speed at 25 FPS with acceptable AP and AR scores.

In the future, it is necessary to improve the performance of the model. Additionally, this model should be further optimized to be faster.

References

[1] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 717–732.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[3] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

TABLE III

SCORES AND SPEEDS OF NETWORKS WITH SEVERAL KINDS OF CONVOLUTION: VANILLA CONVOLUTION, DEPTHWISE SEPARABLE CONVOLUTION, EXPANDED DEPTHWISE SEPARABLE CONVOLUTION FOR THE KEY-POINTS ESTIMATION NETWORK ON COCO DATASET.

Convolution	AP	AP_{50}	AP_{75}	AP_M	AP_L	AR	AR_{50}	AR_{75}	AR_M	AR_L	Speed (FPS)
Vanilla Conv	46.1	73.4	47.9	37.1	59.9	54.8	81.5	56.9	44.1	69.5	11.9
DWConv	37.8	67.3	36.4	28.9	50.9	46.4	75.9	46.2	35.6	61.4	25.2
EDWConv	38.5	67.7	37.9	29.7	51.6	47.0	76.0	47.1	36.0	62.2	25.0

Fig. 2. Qualitative results of the proposed model on validating samples from Coco dataset.

[4] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik, "Articulated pose estimation using discriminative armlet classifiers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3342–3349.

[5] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[8] V.-T. Hoang, V.-D. Hoang, and K.-H. Jo, "An improved method for 3d shape estimation using cascade of neural networks," in *Proceedings of the IEEE International Conference on Industrial Informatics*, 2017, pp. 285–289.

[9] V.-T. Hoang and K.-H. Jo, "3d human pose estimation using cascade of multiple neural networks," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2064–2072, 2019.

[10] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1314–1324.

[11] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, arXiv preprint arXiv:1704.04861.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[13] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 34–50.

[14] 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.

[15] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 740–755.

[16] R. C. Luo and S. Y. Chen, "Human pose estimation in 3-d space using adaptive control law with point-cloud-based limb regression approach," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 1, pp. 51–58, 2016.

[17] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 483–499.

[18] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4903–4911.

[19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of the Neural Information Processing Systems*, 2019.

[20] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 588–595.

[21] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4929–4937.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[25] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3674–3681.

[26] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2820–2828.

[27] T.-J. Yang, A. Howard, B. Chen, X. Zhang, A. Go, M. Sandler, V. Sze, and H. Adam, "Netadapt: Platform-aware neural network adaptation for mobile applications," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 285–300.

[28] J. Yu, C. Hong, Y. Rui, and D. Tao, "Multitask autoencoder model for recovering human poses," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 6, pp. 5060–5068, 2018.

[29] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.