Distilled RSN: Lightweight Pose Estimation Using Knowledge Distillation

Jiu Yi
Graduate School of FSE
Waseda Unversity
Tokyo, Japan
yijiu@fuji.waseda.jp

Haoyuan Liu
Graduate School of FSE
Waseda Unversity
Tokyo, Japan
liuhaoyuan@akane.waseda.jp

Hiroshi Watanabe
Graduate School of FSE
Waseda Unversity
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—Current research on pose estimation often implements repeated functional blocks in model design to improve detection accuracy. Such a strategy results in increased computational complexity and resource consumption, failing to meet real-time inference demand. We discover this efficiency problem by retraining single and multiple functional modules of RSN and then applying the same metrics for evaluation. The inference performance only improves a bit by introducing computational costs several times. To address this issue, we present a novel model simplification strategy, Distilled RSN, which adopts knowledge distillation to refine the redundant RSN blocks into a single, efficient module. Our experiments demonstrate that our method outperforms many lightweight pose approaches in the COCO keypoint dataset. Compared to the method that applies only the original single RSN module for pose inference, we improved accuracy by 1.6% by our strategy.**

***Index Terms*—pose estimation, knowledge distillation**

I. Introduction

Human pose estimation involves the task of determining the spatial coordinates of key points within a human pose image. A significant portion of prior research has focused primarily on enhancing accuracy performance, often neglecting to address efficiency concerns [1], [2]. These methods tend to incorporate multiple repetitions of functional modules within the model, resulting in a significant increase in computational costs. The classical pose estimation methods such as Hourglass, and HRnet, utilize multi-stage network architecture, these methods improve limited performance but introduce an increasing number of parameters. We also experimented to prove the efficiency problem by retraining a single RSN module and multi-stage RSN and applying the same evelation metrics, it showed that the accuracy only improved about 5% but params increased about 4 times.

To address this issue, we propose a new model simplification method that utilizes knowledge distillation to consolidate redundant RSN blocks into a single efficient module. Our method makes full use of a single distilled module and keeps a good balance between accuracy and efficiency. In the end, we compared the method with other lightweight pose estimation approaches. It shows better performance on accuracy and also a smaller model size.

We summarize our contributions as follows:

(i) We identify the efficiency problem inherent in the classical model design, which involves repeated functional modules.
(ii) We propose Distilled RSN, a new model simplification method that enables training more efficient pose estimation networks.

II. Related Work

A. Pose Estimation

Numerous previous studies on pose estimation have focused on enhancing the accuracy performance of models. They are mainly divided into two categories: Regression-based and heatmap-based [1], [2] methods. The heatmap-based method achieves superior inference accuracy by fully utilizing the spatial information from feature maps, compared to the regression-based method. Consequently, it dominates the domain of high-accuracy pose estimation. RSN (Residual Steps Network) is a heatmap-based approach that leverages the Pose Refine Machine to attain top-tier accuracy on the COCO 2019 benchmark. Given the exceptional performance of this method, a key area of improvement is simplifying its multi-stage architecture to reduce the parameter count while preserving efficiency. To this end, our research focuses on optimizing the multiple redundant modules to a single RSN module to achieve maximum inference accuracy while ensuring the model remains computationally efficient. Our experimental approach builds on this foundation, aiming to enhance the model's performance and efficiency through a streamlined design that maintains high accuracy.

B. Knowledge Distillation

Knowledge distillation [3] is widely employed in the field of model compression. Its main idea is to transfer knowledge from a complex model with high accuracy to a compact one, thereby achieving higher efficiency while preserving performance. Knowledge distillation has been employed to enhance model efficiency within the field of pose estimation. For instance, FPD [4] developed a lightweight pose estimation network that effectively balances high accuracy with computational efficiency. Their approach involves using 8 stacked hourglass modules to distill the capabilities of a 4 stacked hourglass network, resulting in a 0.8% accuracy

979-8-3503-5506-2/24/$31.00 ©2024 IEEE

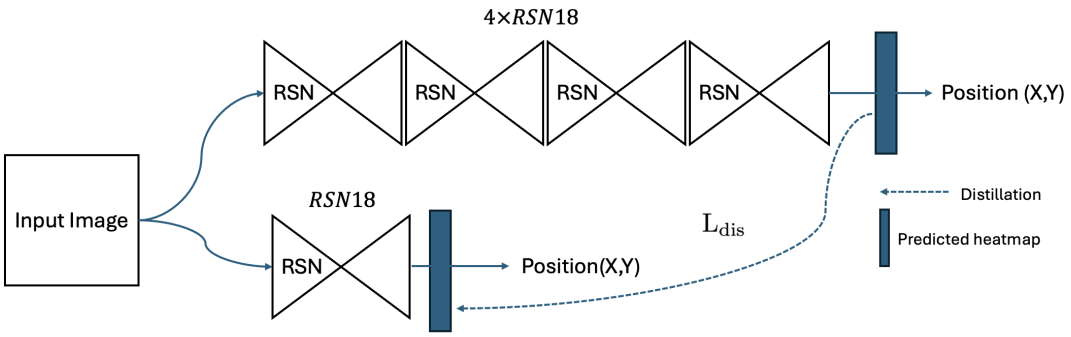

Fig. 1. An overview of the proposed distilled RSN method, 4×RSN18 will be discarded after transferring knowledge and we keep distilled RSN as the final model

improvement over the original 4 hourglass network. Given that RSN achieves top accuracy on the COCO pose benchmark and also utilizes a multi-stage network architecture, applying similar distillation techniques to RSN presents a promising opportunity to enhance its performance while preserving its efficiency. Drawing inspiration from FPD's success, we apply distillation strategies to RSN to create an innovative, efficient pose estimation network that achieves superior accuracy while maintaining computational efficiency.

III. Method

A. The Efficiency Problem

The inefficiency arises from the repeated use of functional modules. Classical models like Hourglass [1] often employ multiple repetitions of these modules, known as intermediate supervision. This design choice allows subsequent modules to re-access high-level features for further processing, thereby enhancing model performance. However, we question the efficiency of this approach and conduct experiments to assess its impact, focusing on the Residual Steps Network [2] (RSN) as an example. RSN, characterized by multiple RSN modules (e.g., 4×RSN), is compared against a single RSN through re-training and evaluation using consistent metrics. Surprisingly, while 4×RSN increased parameters fourfold compared to a single RSN, it only marginally improves accuracy by approximately 5% as shown in Table I. This highlights the inefficiency caused by redundant module repetitions. To address this issue, we employ knowledge distillation to extract a single distilled module from these repeated blocks, eliminating redundancy and enhancing overall model efficiency.

TABLE I
Comparison of 4×RSN18, RSN18, and Distilled RSN18 Model

Method	Average Precision (%)	Parameters (M)
4×RSN18 [2]	73.3	38.46
RSN18 [2]	68.2	9.15
Distilled RSN18	69.8	9.15

B. Distilled RSN

The proposed method is to use repeated 4×RSN18 to distill a single RSN18 module. The knowledge distillation method we applied is called response-based distillation. The response-based distillation, as one of the knowledge distillation methods, utilizes the teacher model's last layer output to supervise the student model. After transferring the knowledge, we discard the original repeated modules and only keep the distilled RSN as the final result. In the distillation part, we choose 4×RSN18 as the teacher model and single RSN18 as the student model.

The original RSN belongs to the heatmap-based pose estimation method. To represent the ground truth joint labels, we generate a heatmap $\mathbf{m}_k$ for each single joint k ($k \in \{1, \cdots, K\}$ by centering a Gaussian kernel around the labeled position $\mathbf{z}_k = (x_k, y_k)$. More specifically, a Gaussian heatmap $\mathbf{m}_k$ for the k-th joint label is written as:

$$\mathbf{m}_k(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{[(x - x_k)^2 + (y - y_k)^2]}{2\sigma^2}\right) \quad (1)$$

where (x, y) specifies a pixel location and the hyper-parameter σ denotes a pre-fixed spatial variance.

The MSE loss function is then obtained as:

$$\mathcal{L}_{\text{mse}} = \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{m}_k - \hat{\mathbf{m}}_k\|_2^2 \quad (2)$$

where $\hat{\mathbf{m}}_k$ refers to the predicted heatmap for the k-th joint.

The overall loss function of our method contains two parts, the distillation loss(representing the loss between the teacher model and student model generated heatmap) and the student loss(representing the loss between student and ground truth heatmap). These losses can be formulated as

$$\mathcal{L}_{\text{dis}} = \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{m}_k^{\text{s}} - \mathbf{m}_k^{\text{t}}\|_2^2 \quad (3)$$

$$\mathcal{L}_{\text{stu}} = \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{m}_k^{\text{s}} - \mathbf{m}_k^{\text{g}}\|_2^2 \quad (4)$$

$$\mathcal{L}_{\text{overall}} = a\mathcal{L}_{\text{dis}} + (1 - a)\mathcal{L}_{\text{stu}} \quad (5)$$

where m_k^{s} and m_k^{t} specify the heatmap for the k-th joint output by the student model (RSN18) and teacher model (4×RSN18),

TABLE II
EVALUATION OF OUR PROPOSED DISTILLED RSN ON COCO TESTING SET

Method	AP	AP@50	AP@75	AP_L	AP_M	AR	Year
Faster R-CNN [5]	64.4	85.7	70.7	69.8	61.8	–	2017
OpenPose [6]	64.2	86.2	70.1	68.8	61	–	2018
Simple Pose [7]	68.1	–	–	70.5	66.8	88.2	2019
Lite-HRNet-18 [8]	66.9	89.4	74.4	72.2	64.0	72.6	2021
Lite-HRNet-30 [8]	69.7	90.7	77.5	75.0	66.9	75.4	2021
Distilled RSN(ours)	69.8	88.8	77.3	76.1	66.6	75.9	2024

and m_k^{g} is the k-th joint heatmap generated from the ground truth image and annotations. a is the balancing weight between the two loss terms.

IV. EXPERIMENT

Dataset. The COCO [9] keypoint dataset presents naturally challenging imagery data with various poses. It contains more than 200k images and 250k person instances labeled with keypoints. In evaluation, we follow the commonly used train/val/test split.

Trainning details. For the experiment of efficiency problem, we train the 4×RSN with 768,000 iterations and specify the batch size as 16. For the single RSN model, we use 460,800 iterations and choose a batch size of 16. For the distilled RSN method, we use the checkpoint of 4×RSN from the efficiency experiment, and then we distill and retrain the RSN by the supervision of the pre-trained 4×RSN model. In the distillation part, for calculating the overall loss. It takes 2 times time and gpu memory in the training process compared with training a single RSN module cause it needs to calculate 2 heatmaps(student and teacher), so it increases the time for training a lot. We set the a as 0.5 to balance the supervision between ground truth and the teacher model. We test the hyperparameter a value from 0.4 to 0.6. Theoretically, the value more closer to 1, and the predicted heatmap would be closer to the teacher model's output. The value closer to 0, our method will be the same as training the student model from scratch. After comparing the final accuracy of different experiments that set a as a different value, 0.5 achieves the best accuracy. As a result, we improved the 1.6% accuracy compared to the original single RSN as shown in Table I.

Results. Table II presents the evaluation results on the COCO dataset, comparing our Distilled DSN model with other pose estimation approaches. The results demonstrate that our Distilled DSN consistently surpasses several classical and lightweight models in terms of accuracy.

A. Comparation To Classical Methods

Specifically, when compared to classical methods such as Faster R-CNN [5] (with 25.6M parameters and 64.4% AP) and OpenPose [6](with 52.3M parameters and 64.2% AP), our Distilled DSN shows superior performance in both accuracy and efficiency. This indicates a notable advancement in both aspects compared to these traditional models.

B. Comparation To Lightweight Methods

In comparison with FDP [4] and Simple Pose [7], our model also achieves superior performance in accuracy and efficiency. However, when compared to Lite-HRNet-18 [8], which has fewer parameters, our method, while demonstrating better accuracy, faces some limitations related to efficiency due to the larger number of parameters in our model.

V. CONCLUSION

In this work, we present the new model simplification method, Distilled RSN. It aims to address the efficiency problem and train a more efficient model with high-accuracy performance. We evaluated our method on the COCO [9] keypoint dataset, demonstrating superior efficiency and accuracy compared to many lightweight pose estimation models. Compared to the original single RSN [2] blocks, our method improves 1.6% accuracy without introducing other computational complexities. Through a series of experiments, we found that our method surpasses other lightweight pose estimation methods, such as Lite-HRNet-18 [8] by about 3%,and also exceeds traditional methods like FasterRCNN [5] and OpenPose [6] by about 5%.

REFERENCES

[1] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in ECCV 2016, pp. 483-499, Oct. 2016.

[2] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, and J. Sun, "Learning delicate local representations for multi-person pose estimation," in ECCV 2020, pp. 455-472, Aug. 2020.

[3] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, Mar. 2015

[4] F. Zhang, X. Zhu, and M. Ye, "Fast human pose estimation," CVPR, pp. 3517-3526, Jun. 2019.

[5] R. Girshick, "Fast R-CNN," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1440-1448, Dec. 2015.

[6] G. Hidalgo Martinez, "OpenPose: Whole-Body Pose Estimation," Ph.D. dissertation, Carnegie Mellon University, 2019.

[7] J. Li, W. Su, and Z. Wang, "Simple Pose: Rethinking and Improving a Bottom-Up Approach for Multi-Person Pose Estimation," in Proceedings of the AAAI, pp. 9831-9838, Apr. 2020.

[8] C. Yu, "Lite-HRNet: A Lightweight High-Resolution Network," in Proceedings of the IEEE/CVF Conference on CVPR, pp. 10440-10450, Jun. 2021.

[9] Lin, T. Y., Maire, M., B Zitnick, "Microsoft COCO: Common Objects in Context," in Proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, pp. 13-28, Sep. 2014.