Lightweight Human Pose Estimation with Enhanced Knowledge Review

Hao Xu[1]
202212490022@nuist.edu.cn
Shengye Yan*[1,2]
shengye.yan@gmail.com
Wei Zheng[2]
wzheng@minieye.cc

[1] Nanjing University of Information Science and Technology, 210044, Nanjing, China
[2] Shenzhen Youjia Innov Tech Co., Ltd., 518000, Shenzhen, China

Abstract

While current state-of-the-art human pose estimation methods have demonstrated remarkable performance, they frequently suffer from a significant parameter and computation overhead, resulting in slow inference speeds. For this issue, we propose a novel approach to knowledge distillation in lightweight human pose estimation. In previous knowledge distillation methods, the strategies in the cross-stage distillation distillation overlooked semantic mismatches caused by the differing complexities of teacher and student networks, potentially leading to negative regularization. To address this issue, we propose a novel method based on the cross-stage knowledge distillation framework. In the cross-stage knowledge distillation process, we transform student features in different stages through multiple receptive field feature transformations, by expanding the receptive fields of student features to better align them to the receptive fields of teacher features. We compute the similarity matrix between student and teacher features. By associating the features of both, we obtain cross-attention weights to facilitate effective cross-layer distillation interaction. At the output stage of the model, we replace the heatmap-based keypoint representation method with a classification coordinate-based approach, reducing the inference memory by 20% and speeding up inference time. Additionally, the vanilla knowledge distillation is performed on the output horizontal and vertical coordinates. Extensive experiments on the MPII and COCO datasets validate the effectiveness of our approach.

1 Introduction

Human Pose Estimation (HPE), as one of the fundamental tasks in computer vision, is widely applied in human behavior recognition, human-object interaction recognition, and human-computer interaction [9, 15, 28, 30]. Existing research shows that high-precision human pose estimation models often require larger backbone networks, more computational resources, and inference memory, making them difficult to deploy on resource-constrained mobile devices. To reduce computational costs, lightweight backbone networks such as MobileNet [12] and ShuffleNet [41] have significantly reduced the model's computational parameter volume, but they have deficiencies in inference accuracy. Therefore, we address this issue by adopting the knowledge distillation approach.

© *Corresponding author.

Knowledge distillation [8] is a technique used to transmit information from a large teacher network to a small student network. Previous methods of human pose estimation distillation generally involve feature distillation at the layer-to-layer and block-to-block levels [10, 25, 39], capturing rich knowledge by distilling specific representations of the generated feature maps to achieve performance improvements. However, the intermediate layers of teacher and student networks often have different model complexities, and distillation strategies in cross-stage may lead to semantic mismatches when distilling features at those layers, thereby negatively affecting the learning process of the student network. In this paper, we further explore the role of knowledge distillation in pose estimation based on the cross-stage knowledge distillation. The experiments are designed by jointly utilizing multi-layer features in the intermediate layers of teacher and student models, as well as expanding the receptive field of the student network . Attention is used to adaptively allocate weights in multi-layer feature distillation to reduce the negative regularization effects caused by different features. In the final part of the network, we replace the Heatmap-based head with a SimDR-based head [16], decoupling coordinate representations to replace heatmap regression coordinates, reducing computational and memory overhead in the upsampling . In the experiment, we Generate the supervised signal using a 1D gaussian distribution. Based on this, vanilla knowledge distillation [11] is applied to the outputs of the teacher and student networks. The contributions of this paper are as follows:

• We introduce a novel technique for cross-stage knowledge distillation in pose estimation, which involves combining multi-layer feature distillation with cross-attention weight allocation and expanding the receptive field of the student feature maps. This approach aims to reduce the negative regularization effects caused by semantic mismatches during distillation.

• In terms of complexity and speed, we achieve state-of-the-art performance through extensive experiments on the COCO and MPII human pose estimation datasets.

2 Related Work

2.1 Lightweight Human Pose Estimation

Lightweight human pose estimation, as a part of HPE is essential in scenarios with resource-constrained devices. One way to achieve model lightweighting is to borrow feature extraction modules from general lightweight models [12, 19, 41]. while maintaining high-resolution feature maps. LiteHRNet [38] introduces shuffle block to lighten the model, Efficient pose [9] introduces a large number of depth separable convolutions in the basic feature extraction module design process. Such methods significantly reduce the number of model parameters and computational complexity. Another way is to abandon keypoint heatmaps. The traditional heatmap-based methods generate Gaussian heatmaps as labels using a 2D Gaussian distribution, requiring multiple upsampling layers to restore feature map resolution from low to high, resulting in greater computational costs. The regression method [2] uses neural networks to directly regress keypoint coordinates. Faster Pose [7] designs an RCE loss function to improve model accuracy. RLE [14] uses maximum likelihood estimation to develop efficient and effective regression-based methods. The SimDR [16] reformulates HPE as two classification tasks for horizontal and vertical coordinates. It can omit additional refinement post-processing and, in certain configurations, eliminate upsampling layers, thus providing a simpler and more efficient approach for HPE. The overall architecture of our network, as show in Fig.2, replacing the heatmap-based head [40] with a SimDR classification head not

only reduces memory overhead and speeds up inference time but also yields improved results by obtaining horizontal and vertical coordinates through vanilla distillation output.

2.2 Knowledge Distillation

Knowledge distillation is a model compression method that does not alter the network structure. Hinton *et al*. [11] first proposed supervising a small student network using soft labels generated by the teacher network's outputs. This method created for classification tasks is called logit-based distillation. This series of works [10, 13, 33, 34] focuses on the knowledge of the final output labels, without the need to consider the internal structure or feature representation of the neural network model, directly utilizing the model's predicted output for samples. Feature-based knowledge distillation refers to using the intermediate features of the teacher network as soft labels. FitNet [25] uses the intermediate layer features of the teacher model's network to supervise the student model and make it fit these intermediate features. Transitioning from logit-based distillation to feature-based distillation, these approaches transfer knowledge from intermediate layers [35, 37] and extend distillation techniques to a variety of tasks including detection [6, 36], segmentation [26], and others.

In recent times, knowledge distillation has been effectively employed in human pose estimation [17, 17, 21]. However, in the process of using a teacher model to guide a student model, inevitable negative regularization effects arise due to feature mismatch when conducting Layer-to-Layer, Block-to-Block distillation with teacher-student networks of different complexities. Therefore, in the feature-level distillation stage of our network, we designed a distillation module based on multi-layer and multi-receptive field guidance.

3 Method

3.1 Pose Estimation Framework with Cross-stage KD

Traditional cross-stage knowledge distillation methods allow student models to learn from multiple stages of the teacher model, thereby enhancing the performance and generalization ability of the student model. However, these methods often fail to fully consider the structural differences between teacher and student models, which may lead to inefficient learning and increased difficulty. With the introduction of the Knowledge Review approach [4], models can balance the learning of shallow and deep knowledge by continuously reviewing shallow knowledge during training, making it more suited to the structure of the student model. Therefore, this paper adopts a framework based on Knowledge Review, introducing cross-stage knowledge distillation, using HRNet as the teacher network and LiteHRNet as the student network, as illustrated in the Fig.1. Based on the cross-stage structure, we employ the multiple reception field transform module(MFT) during the feature distillation phase to effectively expand the receptive field of the output features of the student network. In the multi-layer distillation process, the distillation of shallow features and deep features share the same weights, which may inevitably lead to negative regularization effects as the network depth increases. Therefore, we have designed a cross-attention weight allocation module in the network to dynamically adjust the distillation weights between different stages. We replace the heatmap-based head with a SimDR-based classification head, as shown in Fig.2. Furthermore, by utilizing the horizontal and vertical coordinates of keypoints obtained from classification, we conduct teacher-student network distillation on vanilla knowledge distillation.

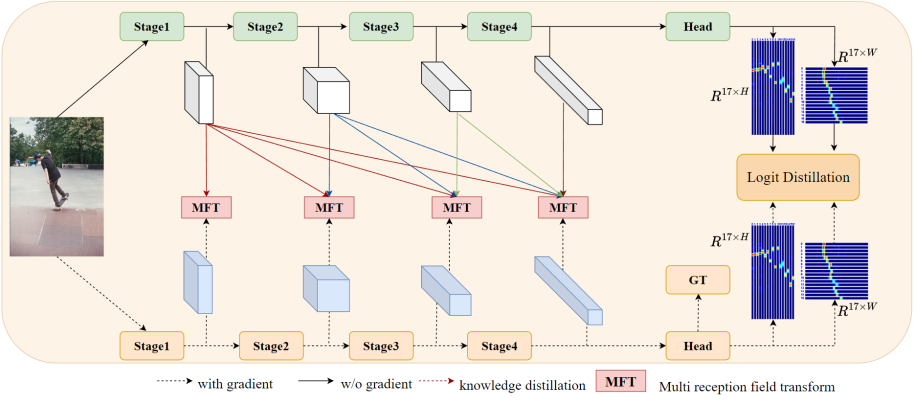

Figure 1: An overview of the proposed Knowledge Distillation framework

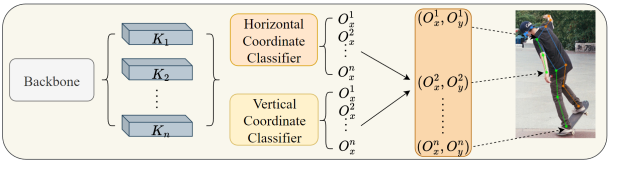

Figure 2: Simple Decoupled coordinate Representation

In our distillation process, each stage of the teacher model is strategically designed to hierarchically guide the corresponding and subsequent stages of the student model. This ensures a comprehensive and gradual transfer of knowledge, enabling the student network to continuously review the shallow features from the teacher throughout the learning process. In general, the loss in the cross-stage distillation is initially expressed as

$$L_{fkd} = \sum_{l=1}^{L} \sum_{m=l}^{M} D(F_{t_l} - f_{align}(F_{s_m})) \tag{1}$$

where F_{t_l} denotes the teacher features from stage l, and F_{s_m} denotes the student features from stage m, where L and M denote the number of stages in the teacher and student networks. We use a transformation f_{align} to match the dimensionality of the student features to that of the teacher features. The Mean Squared Error (MSE) loss is employed as the distance function D between the student features and the teacher features.

3.2 KD with Multiple Reception Fields in HPE

In the process of knowledge distillation, selecting an appropriate feature transformation strategy is crucial for enhancing the performance of the student network. In Review KD, a pooling pyramid module is employed to process the outputs from both the teacher and student networks, segregating the features into different levels of contextual information for loss computation. However, in the pose estimation distillation framework, the teacher network is often more complex than the student network, and larger window pooling operations may

lead to the loss of significant spatial information. Therefore, we have designed a multiple reception field transform module in our network, which incorporates multi-size convolution and pooling operations. This module increases the receptive field of the student features during multi-stage distillation, allowing the student network to perceive contextual information over a larger area and better capture the interrelationships and spatial layouts of various joints in complex poses.

As shown in Fig.3, in the module MFT, features from various stages of the student network are considered. When the dimensions of the output features from different stages of the student network do not align with those of a specific stage of the teacher model, Transposed Convolution is used to adjust the size of the student feature maps. Additionally, 1×1 convolutions are employed to align the channels, thereby facilitating the alignment of student features with teacher features. Distillation is then conducted through branches with multiple reception fields, allowing for effective knowledge transfer from the teacher's features.

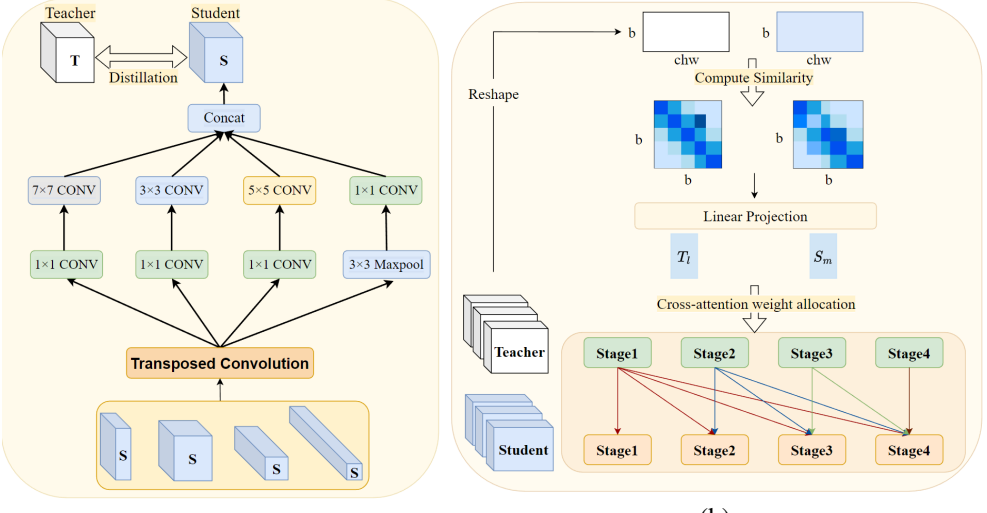

Figure 3: (a) Multiple reception field feature transform(MFT). (b) Weight allocation. Multiple layers' similarity matrices are computed to assign the attention weight.

3.3 KD with Adaptable Cross-stage Learning Weight in HPE

Consequently, in the process of stage-to-stage or layer-to-layer distillation, as the network depth increases, negative regularization effects may occur due to feature mismatch between the corresponding teacher and student networks.

However, in our approach, based on the cross-stage knowledge distillation framework, we dynamically adjust the distillation weights during the teacher-to-student guidance process across different stages through cross-stage weight allocation. The weight allocation module dynamically adjust the cross-stage distillation process by leveraging the similarity matrix [27] between feature maps to adaptively allocate the loss weights $\gamma(s_m, t_l)$ through self-attention, thereby mitigating negative effects.

$$L_{fea} = \sum_{l=1}^{L} \sum_{m=l}^{M} \gamma(s_m, t_l) D(F_{t_l} - f_{align}(F_{s_m})) \tag{2}$$

Where L represents the number of stages in the teacher network, and M represents the number of stages in the student network, $\sum_{l=1}^{L}\sum_{m=1}^{M}\gamma(s_m,t_l)=1$.

In order to address the negative regularization effect arising from feature mismatch between the teacher and student during distillation, we introduce distillation weights for multi-layer distillation by leveraging the similarity between teacher and student feature maps during the associated training process. Since the proximity of pairwise similarity matrices may be considered as a reliable indicator of the underlying semantic similarity [27] during the distillation of feature maps. First, compute the similarity matrix:

$$S_m = F_m^S \cdot (F_m{}^S)^{Transpose}; \quad T_l = F_l^T \cdot (F_l^T)^{Transpose} \tag{3}$$

where $F_m^S \in R^{b\times chw}$ is reshaped from $F_{s_m} \in R^{b\times c\times h\times w}$, and F_l^T is the same.Therefore S_m and T_l are $b\times b$ matrices. Then we project the obtained similarity matrices through two linear layers. Each sample vector from the resulting similarity matrix is processed through a linear layer followed by an activation function, then passed through a normalization layer. The output vectors are standardized using the L2 norm, effectively refining the sample features. Then we obtain the corresponding sample vector $S_m[i]$ and $T_l[i]$. Finally, we calculate attention weights through a similarity feature map associating students and teachers.

$$\gamma^j(t_l,s_m) = \frac{e^{S_m[i]\cdot T_l[i]^{Transpose}}}{\sum_{l=1}^{L}\sum_{m=l}^{M} e^{S_m[i]\cdot T_l[i]^{Transpose}}} \tag{4}$$

Where i denotes the i-th instance in the batch b, $\sum_{i=1}^{b}\sum_{l=1}^{L}\sum_{m=1}^{M}\gamma(s_m,t_l)=1$. By distributing the attention weights, we obtain the final distillation loss.

$$L_{fkd} = \sum_{i}^{b}\sum_{l=1}^{L}\sum_{m=l}^{M}\gamma^j(t_l,s_m)D(F_{t_l}[i]-f_{align}(F_{s_m}[i])) \tag{5}$$

3.4 Vanilla KD for the Output Horizontal and Vertical Coordinates

As shown in Fig. 2, SimDR transforms the feature representation of the backbone network's output stage into a classification task for horizontal and vertical coordinates of keypoints. By training the network to classify coordinates instead of using heatmap-based methods with multiple upsampling and post-processing steps, SimDR reduces computational costs and speeds up inference. Therefore, we replace the heatmap-based feature representations in both the teacher and student networks with SimDR and, based on this, employ the vanilla knowledge distillation [11] for the outputs of the teacher and student. The loss function based on coordinate classification is as follows

$$L_{task} = -\sum_{n=1}^{N}\sum_{k=1}^{K} M_{n,k}\cdot\sum_{i=1}^{L}\frac{1}{L}\cdot V_i log(S_i) \tag{6}$$

where N denotes the number of people in one batch, K denotes the number of keypoints, L denotes the localization bins for vertical and horizontal coordinates. $M_{n,k}$ denotes the weight mask used to mark the unseen keypoints, V_i denotes the label value. We generate the supervised signal using a 1D Gaussian distribution, and employ the cross-entropy as the loss function.

We adopt the SimDR method by replacing the feature representations of the teacher and student. At the output stage of the model, based on vanilla distillation, we use KL divergence as the loss function to supervise the output of the student model.

$$L_{vkd} = -\frac{1}{N} \cdot \sum_{n=1}^{N} \sum_{k=1}^{K} KL(T_i, S_i) \tag{7}$$

where T_i denote the prediction of the teacher, S_i denotes the prediction of the student.

3.5 Overall Loss

In summary, the overall loss during the training process of the student pose estimation network is as follows

$$L = L_{task} + \alpha \cdot L_{fkd} + \beta \cdot L_{vkd} \tag{8}$$

where L_{task} is the original loss for the pose estimation, α and β are the hyperparameters to balance the total loss.

4 Experiments

4.1 Implementation Details

We evaluate our method on the MPII [24] and COCO dataset [18]. The MPII dataset is a large-scale human pose dataset containing 24,920 images collected from YouTube videos that contain human bodies, with 16 key points annotated for each human body. The COCO dataset is a large-scale image dataset, it contains approximately 328,000 images and labels approximately 250,000 human body instances with each human body containing 17 key points. Data augmentation uses random image cropping and random transformation of the bounding box, image flipping with angles ranging from -30° to 30°, and image scaling ranging from 0.75 to 1.25. Our experiments are conducted on two 3080 GPUs using MMPose [6] based on PyTorch [23]. The model optimization uses the Adam optimizer, with an initial learning rate of 5e-4, with the weight decay factor 0.1, a total of 300 training epochs, and a batch size of 32 during training. The hyperparameters of the loss function α = 0.005, β = 0.04. In the experiment, we use HRNet-W32 as the pretrained teacher model and selected the lightweight backbone network LiteHRNet. Additionally, we utilized SimDR as the keypoint representation method in MPII dataset, achieving improvements in reducing Inference Memory and speeding up inference time. In the subsequent distillation strategy, we default to using SimDR as the head of the network.

4.2 Main Result

MPII Dataset: With the SimDR head, improvements have been achieved in scenarios where models of varying complexities are based on reducing model parameters and accelerating inference. Specifically, LiteHRNet18 achieved a 1.6% increase in PCKh@0.5 while reducing inference memory by 20%. In Table 1, we present our model based on this and compare it with state-of-the-art models and lightweight models through multi-stage distillation. Table 2 summarzies results on MPII with the teacher and student using different distillation methods. Our vanilla KD method is conducted on the SimDR feature representation method we propose. In our approach, both vanilla and Feature-based distillation methods are employed.

Distillation Mechanism	Vanilla	Single Layer		Multiple Layers			
Method	KD [10]	FitNet [25]	RKD [27]	AF [39]	SP [27]	ReviewKD [9]	ours
PCKh@0.5	86.74	86.53	86.94	87.12	87.37	87.56	87.73

Table 1: Results on MPII validation set with the teacher HRNet-W32 and student LiteHRNet18 having different architectures.

Model	Input size	Params	Infer memory	Speed	PCKh@0.5
ViTPose-B([34])	256 × 256	86M	14090MB	——	92.3
HRNET48([8])	256 × 256	64M	——	3.14s	91.4
HRNET32([28])	256 × 256	28.5M	8194MB	2.35s	90.8
8-Stack Hourglass([20])	256 × 256	25.1M	5174MB	1.20s	90.2
SimpleBaseline([31])	256 × 256	34.0M	6354MB	1.53s	88.5
MobileNetV2([12])	256 × 256	9.6M	3615MB	1.21s	85.4
ShuffleNetV2([19])	256 × 256	7.6M	1383MB	0.69s	82.8
2-Stack Hourglass([20])	256 × 256	18.6M	——	——	88.6
OKDHP([12])	256 × 256	18.6M	——	——	89.2
LiteHRNet18([33])	256 × 256	1.1M	2900MB	0.25s	86.1
LiteHRNet30([33])	256 × 256	1.8M	3334MB	0.28s	86.9
LiteHRNet18⋆(Multi-KD)	256 × 256	1.1M	2473MB	0.22s	87.7
LiteHRNet30⋆(Multi-KD)	256 × 256	1.7M	2886MB	0.23s	88.9

Table 2: Experimental Results on the MPII validation set. ⋆ means using SimDR as the Head.

COCO Dataset: Table 3 shows the results of our proposed Multi-KD on the COCO test-dev set. Our distillation method further enhances the model's performance compared to the original LiteHRNet, reducing approximately 20% of runtime memory while increasing AP by 1.9% and 2.2%. The above results indicate that our method can still be effectively applied to large-scale datasets.

Model	Input size	Params	Infer memory	Speed	AP
ViTPose-B([34])	256 × 192	86M	15565MB	——	77.1
HRNET48([8])	256 × 192	64M	——	2.94s	74.1
HRNET32([28])	256 × 192	28.5M	8194MB	2.25s	73.4
SimpleBaseline([31])	256 × 192	34.0M	6354MB	1.51s	70.4
MobileNetV2([12])	256 × 192	9.6M	3615MB	1.19s	64.6
ShuffleNetV2([19])	256 × 192	7.6M	1383MB	0.67s	59.9
2-Stack Hourglass([20])	256 × 192	18.6M	——	——	71.7
Lite Pose([25])	256 × 192	1.7M	1174MB	0.31s	40.6
OKDHP([12])	256 × 192	18.6M	——	——	72.8
LiteHRNet18([33])	256 × 192	1.1M	2170MB	0.25s	64.8
LiteHRNet30([33])	256 × 192	1.8M	3334MB	0.28s	67.2
LiteHRNet18⋆(Multi-KD)	256 × 192	1.1M	1794MB	0.22s	66.7
LiteHRNet30⋆(Multi-KD)	256 × 192	1.7M	2816MB	0.24s	69.4

Table 3: Experimental Results on the COCO test-dev set. ⋆ means using SimDR as the Head.

4.3 Ablation Study

In the experiment, we use the hyper-parameters α and β in Equation 8 to balance the training loss. In this section, we conduct the sensitivity study of the hyper-parameters by using HRNet-W32 to distill LiteHRNet18 on MPII dataset, the results are shown in Fig. 5. Finally, $\alpha = 0.005, \beta = 0.04$ is selected. Table 4 shows the ablation experiments conducted on the MPII dataset. In this section, the modules are added one-by-one to measure the efficiency. We use the HRNet-W32 as the teacher. With our proposed method, the result is improved over the baseline. When we further refine the structure with the weight allocation module, the student gains better peoformance. When we finally aggregate the vanilla distillation based on the simDR, the best results are gained.

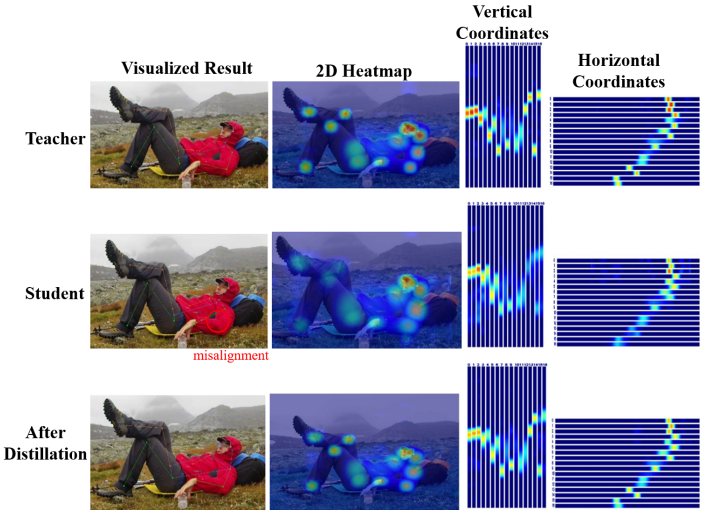

Figure 4: Qualitative comparsions on COCO Dataset. In the visualization of the coordinates, we generate the supervised signal using a 1D Gaussian distribution, and after distillation, a more prominent central Gaussian distribution is produced.

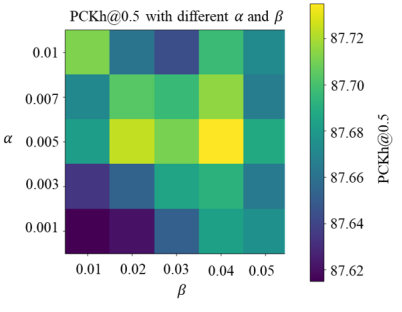

Figure 5: Sensitivity study of hyper-parameters α and β with HRNet-W32 and LiteHRNet18

Cross-stage	MFT	WA	VKD	LiteHRNet18(Multi-KD)	LiteHRNet30(Multi-KD)
✗	✗	✗	✗	85.9	86.7
✓	✗	✗	✗	86.8	87.6
✓	✓	✗	✗	87.3	88.4
✓	✗	✓	✗	87.1	88.1
✓	✓	✓	✗	87.5	88.7
✓	✓	✓	✓	87.7	88.9

Table 4: Ablation result on MPII validation set. MFT: Multi-reception field feature transform (Section 3.2. WA: Weight allocation(Section 3.2). VKD: vanilla knowledge Distillation (Section 3.4)

5 Conclusion

In this work, we design a new knowledge distillation method for lightweight human pose estimation model. During the cross-stage distillation process, we design a multiple reception field feature fusion module to enlarge the student's receptive field for better learning of teacher features. The weight allocation module effectively mitigated the negative regu-

larization effects arising from semantic mismatch in multi-layer distillation processes. In the model's output stage, we replaced the method based on 2D Gaussian heatmap representation with a method based on 1D coordinate decoupling. This not only reduced inference memory and accelerated inference speed but also improved model performance further through vanilla distillation. Multiple proposed models were implemented on COCO and MPII datasets, validating the effectiveness and advancement of the proposed method.

6 Acknowledgement

This work is supported by the Shenzhen Science and Technology Program (Grant No. KQTD20180411143338837).

References

[1] Yanrui Bin, Xuan Cao, Xinya Chen, Yanhao Ge, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Changxin Gao, and Nong Sang. Adversarial semantic data augmentation for human pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 606–622. Springer, 2020.

[2] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 717–732. Springer, 2016.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[4] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021.

[5] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinhong Jiang, and Feng Zhao. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv preprint arXiv:2211.09386*, 2022.

[6] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark, 2020.

[7] Hanbin Dai, Hailin Shi, Wu Liu, Linfang Wang, Yinglu Liu, and Tao Mei. Fasterpose: A faster simple baseline for human pose estimation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(4):1–16, 2022.

[8] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[9] Daniel Groos, Heri Ramampiaro, and Espen AF Ihlen. Efficientpose: Scalable single-person pose estimation. *Applied intelligence*, 51(4):2518–2533, 2021.

[10] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[12] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[13] Tianjin Huang, Lu Yin, Zhenyu Zhang, Li Shen, Meng Fang, Mykola Pechenizkiy, Zhangyang Wang, and Shiwei Liu. Are large kernels better teachers than transformers for convnets? In *International Conference on Machine Learning*, pages 14023–14038. PMLR, 2023.

[14] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11025–11034, 2021.

[15] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 11313–11322, 2021.

[16] Yanjie Li, Sen Yang, Peidong Liu, Shoukui Zhang, Yunxiao Wang, Zhicheng Wang, Wankou Yang, and Shu-Tao Xia. Simcc: A simple coordinate classification perspective for human pose estimation. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022.

[17] Zheng Li, Jingwen Ye, Mingli Song, Ying Huang, and Zhigeng Pan. Online knowledge distillation for efficient pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11740–11750, 2021.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[19] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[20] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016.

[21] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6942–6950, 2019.

[22] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[24] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.

[25] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets. *arXiv e-prints*, art. arXiv:1412.6550, December 2014. doi: 10.48550/arXiv.1412.6550.

[26] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021.

[27] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1365–1374, 2019.

[28] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[29] Yihan Wang, Muyang Li, Han Cai, Wei-Ming Chen, and Song Han. Lite pose: Efficient architecture design for 2d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13126–13136, 2022.

[30] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[31] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[32] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.

[33] Jing Yang, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*, 2020.

[34] Zhendong Yang, Zhe Li, Yuan Gong, Tianke Zhang, Shanshan Lao, Chun Yuan, and Yu Li. Rethinking knowledge distillation via cross-entropy. *arXiv preprint arXiv:2208.10139*, 2022.

[35] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022.

[36] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *European Conference on Computer Vision*, pages 53–69. Springer, 2022.

[37] Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. Vitkd: Practical guidelines for vit feature knowledge distillation. *arXiv preprint arXiv:2209.02432*, 2022.

[38] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-hrnet: A lightweight high-resolution network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10440–10450, 2021.

[39] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[40] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102, 2020.

[41] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.