Cross-Domain Knowledge Distillation for Low-Resolution Human Pose Estimation

Zejun Gu, Zhong-Qiu Zhao, Henghui Ding, Hao Shen, Zhao Zhang, *Senior Member, IEEE*, and De-Shuang Huang, *Fellow, IEEE*

***Abstract*—In practical applications of human pose estimation, low-resolution inputs frequently occur, and existing state-of-the-art models perform poorly with low-resolution images. This work focuses on boosting the performance of low-resolution models by distilling knowledge from a high-resolution model. However, we face the challenge of feature size mismatch and class number mismatch when applying knowledge distillation to networks with different input resolutions. To address this issue, we propose a novel cross-domain knowledge distillation (CDKD) framework. In this framework, we construct a scale-adaptive projector ensemble (SAPE) module to spatially align feature maps between models of varying input resolutions. It adopts a projector ensemble to map low-resolution features into multiple common spaces and adaptively merges them based on multi-scale information to match high-resolution features. Additionally, we construct a cross-class alignment (CCA) module to solve the problem of the mismatch of class numbers. By combining an easy-to-hard training (ETHT) strategy, the CCA module further enhances the distillation performance. The effectiveness and efficiency of our approach are demonstrated by extensive experiments on three common benchmark datasets: MPII, COCO, and Crowdpose. The code is available at https://github.com/guzejungithub/CDKD.**

***Index Terms*—Knowledge distillation, low-resolution image, human pose estimation, cross-domain distillation.**

I. Introduction

HUMAN pose estimation (HPE) is a fundamental task in computer vision which aims to predict the positions of body joints from RGB images [1]–[6]. The recent progress has focused on training methods [7], [8], network structures [9]–[11], and fusion strategies [12], which have notably advanced the accuracy of HPE with high-resolution images [13]–[17].

In real-world application scenarios, images are usually captured in low resolutions, for example, wide-view video surveillance and long-distance shooting. In addition, high-resolution input will bring great computational and memory complexity, which impedes the development of practical applications. However, when current state-of-the-art models are directly applied to low-resolution images, significant performance degradation occurs due to the lack of rich image information. Therefore, it is a critical yet more challenging problem to upgrade the performance of a low-resolution HPE model. One possible way to solve the problem is to transfer the knowledge from high-res models to low-res models.

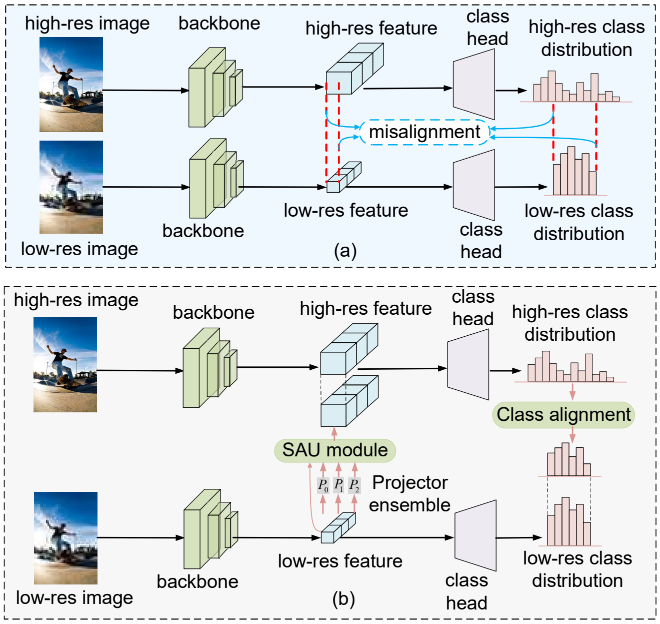

Fig. 1: Comparison between traditional distillation and our CDKD. **(a)** Illustration of the issue of traditional methods for distillation between different resolution models. When distilling knowledge from high-res models to low-res models, there is a problem of misalignment between features and output classes. **(b)** The overview of our proposed CDKD framework.

However, the different input resolutions of the teacher model and the student model lead to the following two problems: 1) The teacher model and the student model do not share the same feature spatial size at the same network stages. In this situation, traditional feature distillation methods can not work. 2) The number of output categories of the teacher model and the student model are not the same. Currently, there are ***NO*** logit distillation studies based on different class numbers in detection tasks, so the logit information in teacher models can not be effectively distilled.

Due to the mismatch in the feature spatial domain and output distribution domain between the teacher model and the student model, we propose a novel knowledge distillation framework termed cross-domain knowledge distillation

Zejun Gu, Zhong-Qiu Zhao, Hao Shen, and Zhao Zhang are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China. (e-mail: guzejunmail@gmail.com, haoshenhs@gmail.com; cszzhang@gmail.com; Corresponding author: Zhong-Qiu Zhao)

Henghui Ding is with the School of Computer Science, Fudan University, Shanghai 200433, China (e-mail: henghui.ding@gmail.com)

De-Shuang Huang is with the Institute of Machine Learning and Systems Biology, Eastern Institute of Technology (e-mail: huangdeshuang@163.com)

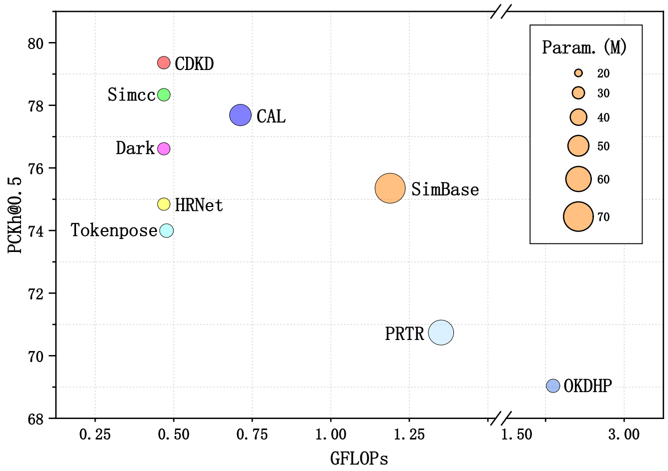

Fig. 2: Comparisons between the SOTA methods and the proposed CDKD on the MPII [18] val dataset with the input resolution of 64×64. Red circles at the upper left corner denote CDKD. It outperforms SOTA models in terms of accuracy (AP), Parameter, and computational cost (GFLOPs).

(CDKD) to resolve it. It mainly includes the following two components: 1) A scale-adaptive projector ensemble (SAPE) module for feature distillation. 2) A cross-class alignment (CCA) module for logit distillation.

The SAPE module resolves the feature size mismatch between the high-res teacher and the low-res student. Firstly, we design a projector to map features into a common space for matching. Feature distillation can be regarded as a multi-task learning process, including feature learning for the original task and feature matching for distillation [19]. In this scenario, the student network might overfit the teacher's feature distributions, leading to less distinguishable generated features for HPE. Adding a projector for distillation would alleviate the overfitting issue. We propose to ensemble projectors for further improvement. According to the theory about ensemble learning [20], different initialized projectors produce different transformed features. The use of multiple projectors helps to improve the generalization ability of the student. To address the body scale variation problem in the natural scene, we propose a scale-adaptive unit (SAU) to capture the multi-scale information. It assigns different weights to each projector by multiple parallel transformations with different receptive fields. Then, it merges the output features of each projector with assigned weights.

The CCA module solves the problem of class number mismatch. As shown in Fig. 1 (b), the probability distribution of the classification output for different resolution models has the same mathematical significance. It can be shared directly between the teacher model and the student model. Meanwhile, the probability values of multiple adjacent classes can be merged by adding them together. In addition, the probability distribution of the classification output is also a fundamental element in calculating distillation loss. Based on this observation, for the high-res teacher, we merge the probability values of adjacent categories to make them consistent with the categories of the student model. It allows for the implementation of logit distillation. Finally, to enhance the effectiveness of distillation, we further propose an easy-to-hard training (ETHT) strategy. It improves distillation performance through curriculum learning.

In summary, our CDKD achieves optimal performance with minimal computational cost. As shown in Fig. 2, CDKD outperforms previous SOTA methods, such as Simcc [21] and CAL [22] with the fewest parameters and GFLOPs.

Our contributions are summarized as follows:

- We propose a novel scale-adaptive projector ensemble (SAPE) module that solves the problem of feature size mismatch between the high-res teacher and the low-res student and attains excellent distillation performance.
- We construct a cross-class alignment (CCA) module to tackle the problem of inconsistent class numbers in logit distillation. Combined with the easy-to-hard training (ETHT) strategy, it further improves the distillation effectiveness.
- Our proposed CDKD is a universal framework that can address the issue of mismatched feature spatial domain and output distribution domain in different distillation tasks. Extensive experiment results demonstrate its efficiency, effectiveness, and universality. It achieves SOTA performance in low-res HPE on MPII [18], COCO [23], and CrowdPose [24] with ***NO*** computational cost increment. When applied to different HPE models and low-res inputs of various sizes, it consistently achieves superior performance.

II. Related Work

A. Human Pose Estimation

The current research on HPE primarily focuses on two aspects: keypoint coordinate representation and applications under complex conditions.

For keypoint coordinate representation, recent works are mainly divided into three mainstreams: regression-based methods [25]–[30], heatmap-based methods [11], [31]–[36], and new keypoint representation methods [1], [21]. Regression-based methods directly regress the coordinates of keypoints within a lightweight framework. Deeppose [30] is the first to propose direct regression of joint coordinates. CenterNet [37] presents a method to accomplish multi-person pose estimation within a one-stage object detection framework, directly regressing joint coordinates instead of bounding boxes. SPM [38] introduces root joints for distinguishing among different person instances, by employing hierarchical rooted representations of human body joints to improve the prediction of long-range displacements for specific joints. The residual log-likelihood (RLE) [25] utilizes normalizing flows to capture the underlying output distribution, which enables regression-based methods to achieve accuracy comparable to SOTA heatmap-based methods. MDN [39] proposes a mixture density network for regression. Regression-based methods have significant advantages in speed, but their accuracy is insufficient.

The heatmap-based methods adopt a two-dimensional Gaussian distribution to represent joint coordinates. Some studies optimize backbones to extract better features. Sun *et al.* [9] introduce a groundbreaking network designed to preserve high-resolution representations throughout the entirety of the process, resulting in substantial performance enhancements. Other studies aim to improve prediction accuracy by reducing quantization errors. Zhang al. [22] propose using a distribution-aware coordinate representation for post-processing of prediction results to reduce quantization errors. Huang *et al.* [40] design a plug-and-play unbiased data processing method that effectively enhances the performance of different models without increasing computational complexity. Wang *et al.* [41] conceptualize the backbone network as a degradation process and recast heatmap prediction as a super-resolution task, which effectively reduces quantization errors and enhances the precision of the model's predictions. Some methods [4], [11] improve their performance by leveraging transformers because they can capture long-range information. Heatmap-based methods are far ahead in terms of performance, but they have the disadvantage of exceptionally high computational cost and slow preprocessing operations. DistilPose [2] combines heatmap and regression methods through knowledge distillation, achieving high accuracy at low computational costs.

New keypoint representation methods explore novel approaches of keypoint representation to reduce quantization errors and leverage keypoint relationships. PCT [1] represents a pose by discrete tokens to model the dependency between the body joints. Simcc [21] effectively minimizes quantization error by transforming the keypoint regression task into a classification problem. They all open up new perspectives on keypoint representation methods.

In addition, many studies are beginning to focus on applications under complex conditions. Yang *et al.* [42] introduce ED-Pose, which achieves fast, concise, and accurate end-to-end HPE. Lee *et al.* [43] present ExLPose to tackle the problem of current models failing to work properly under low-light conditions. Ju *et al.* [44] propose a dataset for human pose estimation in artwork, aiming to apply the HPE model to virtual scenes. Yang *et al.* [45] introduce an efficient full-body pose estimation method (DWPose), aiming at better application in human-computer interaction. Sun *et al.* [46] present the first all-in-one-stage model for expressive human pose and shape estimation, demonstrating outstanding performance. Cai *et al.* [47] propose the first generalist foundation model for expressive human pose and shape estimation, named SMPLer-X. The model demonstrates excellent transferability and achieves superior performance across diverse environments.

B. Low-Resolution Vision Tasks

Wang *et al.* [22] propose a novel confidence-aware learning (CAL) method to reduce quantization errors, thereby improving the performance of HPE models under low-resolution conditions. Li *et al.* [21] design a simple coordinate classification method, which achieves excellent results in low-resolution pose estimation by transforming coordinate regression into a classification problem. Kumar *et al.* [48] present a semi-supervised approach to predict landmarks on low-resolution images by learning them from labeled high-resolution images. It can improve performance on the critical task of face verification in low-resolution images. Chai *et al.* [49] introduce a recognizability embedding enhancement approach to address the very low-resolution face recognition (VLRFR) challenge. Sunkara *et al.* [50] design a new CNN building block (SPD) for low-resolution images and small objects. This block is universal and can replace the downsampling layers in different CNN models, thereby improving the performance of various low-resolution tasks. Xu *et al.* [51] propose a resolution-aware neural network for HPE which can deal with different resolution images with a single model.

C. Knowledge Distillation

Knowledge distillation aims to transfer knowledge from a trained teacher model to a compact and lightweight student model. In recent years, various methods have been proposed for knowledge distillation [45], [52], [53]. These methods fall into two lines of work: 1) feature distillation methods [54]. 2) logit distillation methods [55]. Feature distillation methods distill knowledge using intermediate features, while logit distillation methods are designed to perform distillation on output logits.

Logit distillation was originally proposed using the KL divergence [56], and it has been extended using spherical normalization [57], label decoupling [55], and probability reweighting [58]. MLD [59] introduces a multi-level prediction alignment framework to logit distillation. Through this framework, the student model learns instance prediction, input correlation, and category correlation simultaneously. LD [60] presents a novel localization distillation (LD) method that can efficiently transfer the localization knowledge from the teacher to the student.

While logit distillation methods may seem straightforward and versatile for application across various scenarios, their performance frequently falls short compared to feature distillation. Compared to logit distillation, feature distillation methods are more likely to achieve high performance because they absorb rich knowledge from the teacher model. The work [61] adopts a novel orthogonal projection layer to maximize the distilled knowledge to the student backbone, thereby achieving outstanding performance. CrossKD [62] transfers the intermediate features from the student's head to that of the teacher, generating cross-head predictions for distillation. This approach efficiently mitigates conflicts between supervised and distillation targets. Some methods [63], [64] mitigate the differences in features between the teacher and student models, thus compelling the student model to emulate the teacher model at the feature level. Other methods [65], [66] transmit teacher knowledge by extracting input correlations.

In contrast to existing KD methods that focus on improving the performance of lightweight networks, this paper aims to improve the low-res model by distilling from the high-res model. Recently, a succession of relevant studies has emerged. The work [67] adopts a multi-scale aligned distillation method to tackle the output size mismatch, but this method relies on the feature pyramid (FPN) structure [68] as its backbone. The distillation method improves the low-res model by using

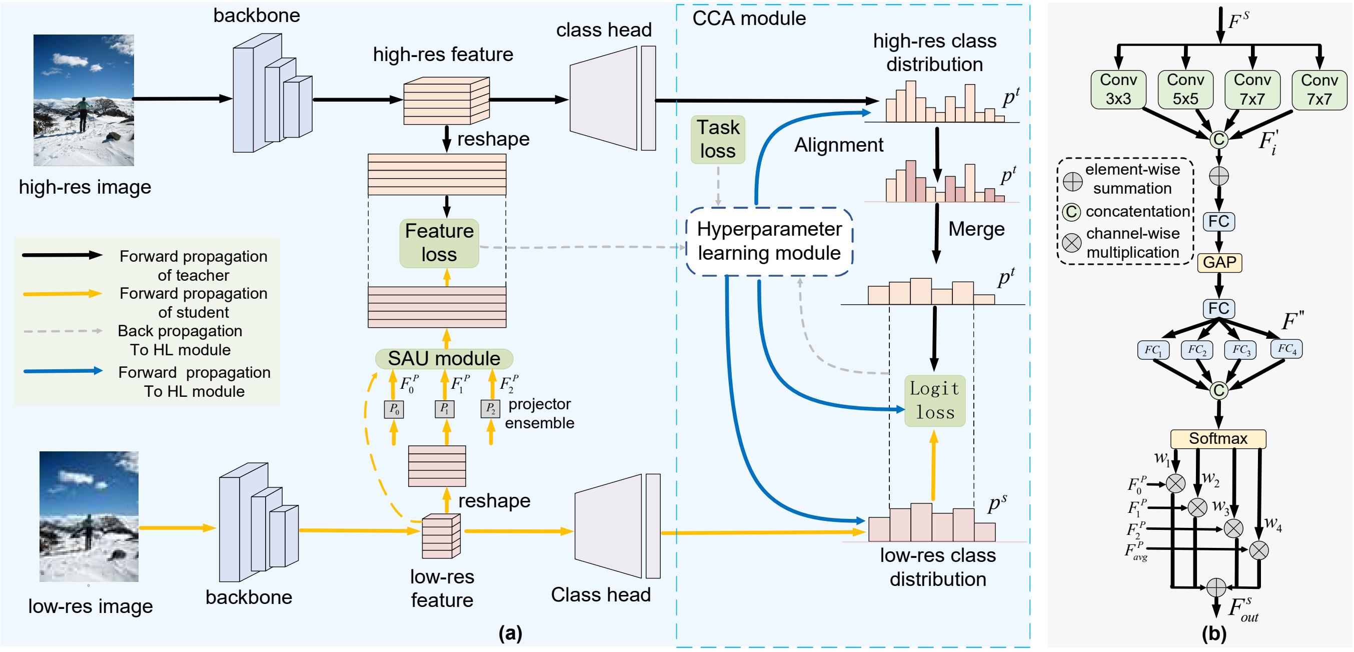

Fig. 3: **(a)** Overall architecture of our proposed CDRD. During training, a well-trained and fixed high-res teacher provides rich knowledge to help the training of a low-res student based on the alignment of features and classes. **(b)** The illustration of our SAU module.

upsampling in the initial stage [69]. However, upsampling introduces noise and incurs substantial computational costs. FMD [70] solves the problem of inconsistent feature sizes, but it can only be applied under the premise that the number of feature channels between the teacher and student is different. FITNETS [71] utilizes convolutional downsampling to align high-res features with their low-res counterparts. Nonetheless, downsampling drastically reduces the rich efficient information hidden in large feature maps. ScaleKD [52] designs a mapping function to align the size of the feature map in different models, but it does not achieve outstanding performance. Our proposed method (CDKD) can solve the above problems and thereby achieve more competitive distillation effects.

III. Method

In this section, we propose a distillation-based human pose estimation framework, cross-domain knowledge distillation (CDKD), as shown in Fig. 3 (a). In CDKD, the teacher is a high-res model, and the student is a low-res model. We transfer the rich details and texture knowledge from the teacher model to the student model during training.

A. Scale-Adaptive Projector Ensemble

The Scale-Adaptive Projector Ensemble (SAPE) aims to align the feature size of the teacher and the student, which consists of a projector ensemble and a scale-adaptive unit.

1) Projector Ensemble : As shown in Fig. 3 (a), our framework is based on the Simcc algorithm. Recent research indicates that feature distillation is more likely to achieve superior performance compared to logit distillation [59]. The last feature of networks is better suited for distillation [72]. One possible reason is that the last feature is closer to the classifier and will directly impact classification effectiveness. Therefore, we adopt the last feature to distill.

We represent the last teacher feature as $F^t \in \mathbb{R}^{B\times C\times H'\times W'}$, where B, C, H', and W' are the batch size, the number of channels, the height, and the width of the last teacher feature maps, respectively. The corresponding student feature is represented by $F^s \in \mathbb{R}^{B\times C\times H\times W}$, where B, C, H, and W are the batch size, the number of channels, the height, and the width of the last student feature maps, respectively. We define m as the scale factor between the high and low resolution, where $H = H'/m$ and $W = W'/m$.

In classification tasks based on distillation, the training process of the student network can be considered as multi-task learning within the same feature space. Therefore, student features are prone to overfit teacher features and would be less discriminative for classification. We add a projector P_i to disentangle the two tasks to improve the student's performance and to match the size of F^t (see Fig. 3 (a)). In addition, we use an ensemble of projectors for further improvement which consists of a group of parallel projectors. Each projector contains a FC layer and a ReLU function.

There are two motivations for adopting ensemble projectors. 1) Projectors with diverse initializations yield distinct transformed features, contributing to the generalizability of the student. 2) The projected student features may include zeros due to the use of the ReLU function in the projector. In contrast, teacher features are less likely to be zeros since the average pooling operation is widely employed in CNNs. The feature distribution gap between the teacher model and the student model is large when using a single projector. In this paper, we use ensemble learning to reduce the feature distribution gap and achieve better generalizability.

2) Scale-Adaptive Unit : The previous work [19] uses a simple additive approach to fuse output features from different projectors. The simple additive approach not only fails to capture multi-scale information in real-world scenarios but also lacks the weight assignment based on the importance of different projectors. Inspired by [73], we propose the scale-adaptive unit (SAU) to deal with the problem. The SAU learns to combine all outputs from different projectors to produce a rich output feature. It is composed of multiple parallel transformations with different receptive fields, exploiting local and global information to obtain a strong low-res feature.

The architecture of the proposed SAU is depicted in Fig. 3 (b). We take the output of each projector and the low-res feature as input of SAU. Initially, we operate it through convolutions with different kernel sizes:

$$F_i^{'} = Conv_i(F^s) \tag{1}$$

where $Conv_i(\cdot)$ denotes convolution operation with different kernel sizes, F^s is the last student feature, $F_i^{'}$ is output of each convolution operation and i is the index of each branch.

Convolutions with different receptive fields help the model to obtain human body information at different scales. Because low-res images contain less useful information, the feature information under a small receptive field does not make sufficient sense. Therefore, we moderately increase the proportion of convolutions with a large receptive field to obtain as much useful information as possible, as shown in Table VI. Next, we fed $F_i^{'}$ to fusion operation:

$$F'' = Fus(Concat[F_1^{'}, F_2^{'}, F_3^{'}, F_4^{'}]) \tag{2}$$

where $Concat[\cdot]$ and $Fus(\cdot)$ denote the concatenation operation and the fusion operation, respectively. This fusion operation is sequentially composed of a fully connected layer, a global pooling layer, and a summation layer. It can adaptively adjust receptive field sizes according to the image content. Then, F'' is applied to select operation:

$$w_k = Select(F^{''}) \tag{3}$$

where $Select(\cdot)$ is the select operation and w_k is the weight value corresponding to each projector output. The select operation is sequentially composed of multiple fully connected layers, a concatenation layer, and a softmax layer. This operation can adaptively select different spatial scales of information with soft attention across channels. Finally, we fuse outputs from multiple projectors via an element-wise summation to obtain the output feature F_{out}^s:

$$F_{out}^s = \sum_{k=1}^{K} w_k \otimes F_k^P \tag{4}$$

where K is the number of branches, F_k^P is the output of each projector, w_k is the corresponding weight value, and F_{out}^s is the final output of this module.

After aligning features from multi-scale spaces, we can use cosine distance between F_{out}^s and F^t to calculate the feature distillation loss:

$$L_{fea} = 1 - \frac{F_{out}^s}{\|F_{out}^s\|_2} \cdot \frac{F^t}{\|F^t\|_2} \tag{5}$$

where $||\cdot||_2$ denotes L2-norm.

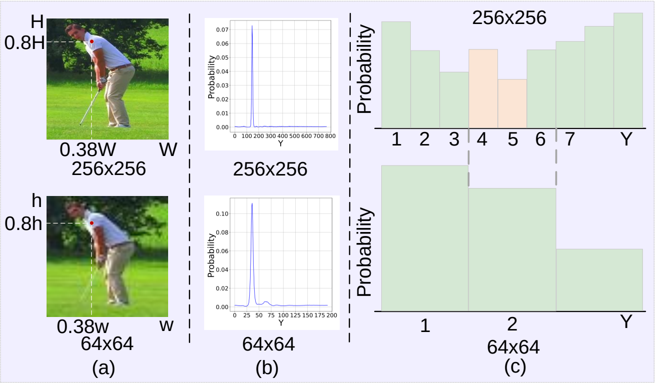

Fig. 4: Illustration of the principle of cross-class distillation. **(a)** In both the 64×64 and 256×256 resolution images, the left shoulder of the person appears at the same normalized coordinates, (0.38, 0.8). **(b)** The model output distribution corresponding to the left shoulder in (a). Each pixel is divided into 3 bins. The "Y" indicates the bin index along the vertical (height) axis. **(c)** A schematic illustration of the local region in (b).

B. Cross-Class Alignment

Another way to transfer knowledge from teacher to student is logit distillation. The feature distillation focuses on optimizing the encoder and does not directly impact the classification head, so we propose a cross-class alignment (CCA) module to refine the student's classification head straightforwardly.

1) Class Alignment : As is shown in Fig. 1 (a), the number of Simcc's output categories is proportional to the input resolution typically. The number of output categories of low-res model and high-res model is different when their resolution is different. However, the current logit distillation is based on the consistency of the number of categories. In our work, we propose a cross-class (CCA) alignment module to resolve this issue. Simcc's classification method can be regarded as one of ordinal regression [74]. There is a certain ordinal relationship among the categories in ordinal regression. Any number of adjacent classes can form an entire meaningful interval. We are interested in exploring the merging of multiple discrete categories of the student to align with the teacher. However, the values of the logit or the feature can not be added together, since they do not have the same computational unit and interpretable real-world meaning.

Combining the characteristics of ordinal regression and the properties of probability distributions, we have the following findings: 1) Regardless of the image resolution, the relative positions of keypoints within the image remain consistent and are independent of the model, as shown in Fig. 4(a). 2) After the final softmax operation, the output probability distributions of models with different resolutions share the same units and practical significance, representing the probability of keypoints at each relative position [21], [75], as illustrated in Fig. 4(b). Hence, these probability distributions can be shared between models of different resolutions (Fig. 4(c)). 3) The values in

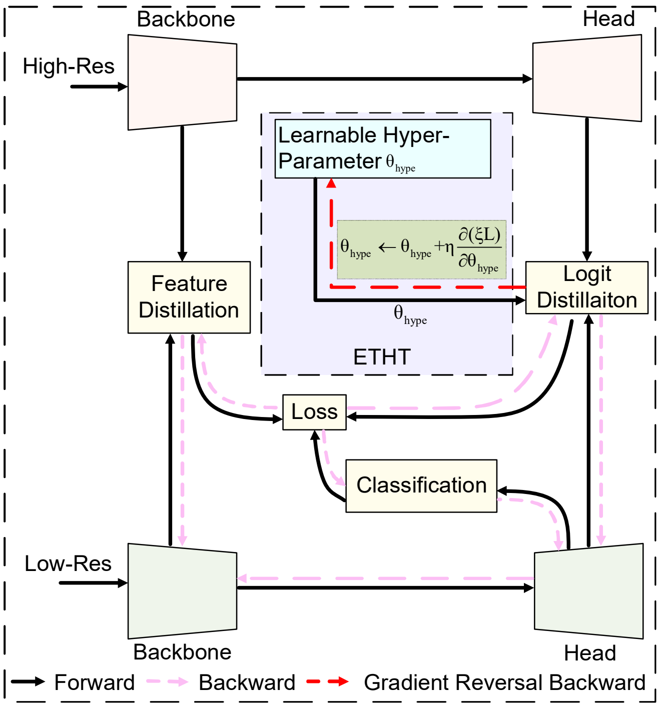
Backbone
Head
High-Res
Learnable Hyper-Parameter θ_{hype}
$\theta_{hype} \leftarrow \theta_{hype} + \eta \frac{\partial(\xi L)}{\partial \theta_{hype}}$
Feature Distillation
Logit Distillaiton
θ_{hype}
ETHT
Loss
Classification
Low-Res
Backbone
Head
Forward
Backward
Gradient Reversal Backward

Fig. 5: The illustration of our proposed easy-to-hard training (ETHT) strategy. **The white-filled box** illustrates the training iteration of CDKD. We integrate ETHT **(purple-filled box)** into the process. Learnable hyperparameters are introduced into the loss function computation. During backpropagation, we adopt an easy-to-hard gradient reversal strategy, gradually increasing the gradient coefficient ξ for backpropagation.

these probability distributions are additive [76]. As shown in Fig. 4(c), the probability of the keypoint falling within the interval [4,5] is the sum of the probabilities of the keypoint being at position 4 and position 5. 4) The probability distribution is also the object of the traditional logit distillation loss [56], serving as a bridge between the teacher model and the student model, which enables it to act as an entry point for cross-category class alignment.

Therefore, we adopt the cross-class alignment method to align the categories of different resolution models. As shown in Fig. 3(a), we sum up the output probability values for every m adjacent categories in the high-res teacher model:

$$\boldsymbol{p}_j^t = \sum_{i=j\times m+1}^{j\times m+m} \boldsymbol{p}_i^t \tag{6}$$

where m represents the ratio between high resolution and low resolution, $\boldsymbol{p}_i^t$ denotes the output probability values of i-th category in the high-res model, $\boldsymbol{p}_j^t$ is the j-th probability values after mergers. Then, the teacher model and the student model achieves class alignment, and logit distillation can be performed between them. Based on the above results, we can calculate the logit distillation loss:

$$L_{logit}(p^t, p^s, \tau) = \sum_{j=1}^{J} \tau^2 KL(p_j^t, p_j^s) \tag{7}$$

where τ is the temperature, $\boldsymbol{p}^t$ and $\boldsymbol{p}^s$ denote the output probability distribution produced by teacher and student, $KL(\cdot)$ is the Kullback-Leibler Divergence. With feature distillation loss L_{fea} and logit distillation loss L_{logit}, we can train the student with the total loss as:

$$L_{total} = L_{ori} + \alpha L_{fea} + \beta L_{logit} \tag{8}$$

where L_{ori} is the original task loss for HPE, α and β are the hyperparameters to balance the loss.

2) *Easy-to-Hard Training Strategy:*

In human education, teachers often employ a curriculum strategy of imparting knowledge from simple to challenging, ensuring that students achieve optimal learning outcomes. Influenced by this, many researchers adopt the classic curriculum strategy training the models in an easy-to-hard method, which improves the performance of the models.

Inspired by CTKD [78], we adopt an easy-to-hard training (ETHT) strategy to increase the difficulty gradually. As shown in Fig. 5, this strategy is implemented by adding a learnable hyperparameter learning module to the student model. In this work, We choose temperature τ as the hyperparameter to be trained, which can also be extended to other hyperparameters, such as loss coefficient α and β (see Table VII).

Firstly, this module is optimized in the opposite direction of the student's training, intending to maximize the distillation loss between the student and teacher. It can enhance the training performance of the model by producing adversarial effects.

We apply the gradient reversal method to implement the adversarial process. The traditional gradient descent process is as follows:

$$\theta \leftarrow \theta - \eta \frac{\partial L}{\partial \theta} \tag{9}$$

where θ is the training parameters, η is the is the learning rate, and L is the loss. Our proposed gradient reversal method is as follows:

$$\theta_{hype} \leftarrow \theta_{hype} + \eta \frac{\partial L}{\partial \theta_{hype}} \tag{10}$$

where θ_{hype} is the training hyperparameters. We optimize the hyperparameter module in the opposite direction of gradient descent, which essentially prevents loss from declining and increases the difficulty of model training. In addition, to gradually increase the training difficulty, we introduce a dynamic coefficient ξ in the above process of backpropagation.

$$\theta_{hype} \leftarrow \theta_{hype} + \eta \frac{\partial(\xi L)}{\partial \theta_{hype}} \tag{11}$$

$$\xi = func(\frac{T_i}{T_{\max}}) \tag{12}$$

where $func(\cdot)$ denotes the monotonically increasing function, T_i is current epoch, and T_{max} is total epochs. As training progresses, ξ gradually increases, and the training difficulty gradually increases. Therefore, by adopting this approach, the training performance is effectively optimized.

TABLE I: Comparison with SOTA methods on the MPII validation dataset.

Method	Input Size	Backbone	Head↑	Sho.↑	Elb↑	Wri.↑	Hip↑	Knee↑	Ank.↑	PCKh@0.5↑	PCKh@0.1↑	#Params↓	GFLOPs↓
SimBase [33]	32×32	ResNet-152	40.9	48.2	34.0	21.7	44.6	31.4	22.7	36.8	2.4	68.6M	0.3
SimBase [33]		ResNet-50	41.5	49.5	34.8	22.4	45.8	32.5	23.4	37.8	2.5	34.0M	0.2
HRNet [9]		HRNet-W32	46.0	52.4	40.2	28.4	48.8	37.2	27.5	42.2	2.3	28.5M	0.1
Dark [36]		HRNet-W32	39.0	61.8	46.2	31.9	61.3	45.2	28.8	48.1	5.2	28.5M	0.1
CAL [22]		HRNet-W32	77.1	68.6	48.2	33.1	63.2	46.3	40.6	55.4	7.4	49.3M	0.2
Tokenpose [11]		HRNet-W48	40.2	64.6	49.2	33.9	62.6	46.2	28.9	49.8	5.8	33.1M	0.2
PRTR [26]		HRNet-W32	0.3	1.5	8.1	9.1	16.0	1.6	0.2	5.6	0.3	57.2M	1.5
Simcc (*baseline*) [21]		HRNet-W32	81.3	73.3	54.6	38.4	63.7	50.5	45.6	59.5	7.6	28.5M	0.1
CDKD (*ours*)		HRNet-W32	**82.6**	**74.4**	**56.3**	**40.7**	**64.6**	**52.8**	**48.3**	**61.3**$_{+1.8}$	7.8	**28.5M**	**0.1**
SimBase [33]	64×64	ResNet-152	89.9	85.3	73.9	64.1	76.2	68.6	62.7	75.4	10.6	68.6M	1.3
OKDHP [77]		4-Stack HG	85.6	80.8	66.1	54.3	71.8	61.0	54.4	69.0	8.8	30.9M	2.9
HRNet [9]		HRNet-W32	90.1	85.8	73.0	63.5	75.1	67.3	61.9	74.8	8.6	28.5M	0.6
Dark [36]		HRNet-W32	89.5	87.9	74.6	63.2	78.8	69.4	62.5	76.3	17.9	28.5M	0.6
CAL [22]		HRNet-W32	92.5	88.7	75.8	65.0	80.1	70.2	65.4	77.7	19.5	49.3M	0.8
Tokenpose [11]		HRNet-W48	89.4	87.7	74.1	62.1	78.3	67.0	59.9	75.3	17.0	68.2M	1.2
PRTR [26]		HRNet-W32	90.0	84.2	67.3	52.7	75.9	61.9	53.3	70.5	11.5	57.2M	1.5
Simcc (*baseline*) [21]		HRNet-W32	93.1	88.8	77.3	67.6	79.1	70.4	66.1	78.3	18.2	28.6M	0.6
CDKD (*ours*)		HRNet-W32	**93.6**	**89.8**	**78.4**	**68.1**	**80.8**	**72.0**	**66.4**	**79.4**$_{+1.1}$	18.6	28.6M	**0.6**

IV. Experiments

In this section, we first evaluate the proposed distillation framework on three common benchmark datasets: MPII [18], COCO [23], and Crowdpose [24] . Then we carry out a series of ablation studies to prove the effectiveness of each component. In addition, we show the universality of our proposed framework in various human pose estimation models. Finally, we present sufficient visualization results and perform detailed analyses of the model's complexity and failure cases.

A. Implementation Details

1) Datasets: The MPII dataset includes approximately 25K images containing over 40K subjects with annotated body joints, where 29K subjects are used for training and 11K subjects are used for testing. We adopt the same train/valid/test split as in [79]. Each person instance has 16 labeled joints.

The COCO dataset contains over 200k images and 250k human instances. Each human instance is labeled with K = 17 keypoints representing a human pose. Our models are trained on COCO train2017 with 57k images and evaluated on COCO val2017, which contain 5k images.

The CrowdPose dataset is more challenging than the COCO keypoint dataset, as it contains a large number of crowded and occluded scenes. It consists of approximately 20,000 images and 80,000 human instances, each annotated with 14 keypoints. The dataset is divided into around 10,000 training images, 2,000 validation images, and 8,000 testing images.

2) Evaluation Metric: For the MPII dataset, we adopt the standard Percentage of Correct Keypoint (PCK) metric, which measures the proportion of predicted keypoints that fall within a normalized distance from the ground truth. Specifically, PCKh@0.5 indicates a threshold set to 50% of the head diameter, while PCKh@0.1 corresponds to a more stringent threshold of 10% of the head diameter.

For the COCO dataset, we employ the standard Average Precision (AP) as the evaluation metric, which is computed based on Object Keypoint Similarity (OKS):

$$OKS = \frac{\sum_i exp(-d_i^2/2s^2j_i^2)\sigma(v_i > 0)}{\sum_i \sigma(v_i > 0)} \tag{13}$$

where d_i represents the Euclidean distance between the i-th predicted keypoint and its corresponding ground-truth location, j_i is a per-keypoint constant, v_i denotes the visibility flag, σ is the indicator function, and s indicates the object scale. We report standard average precision and recall scores [23]: AP^{50} (average precision at OKS = 0.50), AP^{75} (average precision at OKS = 0.75), AP (the mean of average precision scores at 10 positions, OKS = 0.50, 0.55,..., 0.90, 0.95), AP^{M} for medium objects, AP^{L} for large objects, and AR (average recall) at OKS = 0.50, 0.55, ... , 0.90, 0.95.

The evaluation of CrowdPose follows the standard protocol of the COCO dataset and uses the Average Precision (AP) based on Object Keypoint Similarity (OKS) as the metric.

3) Training Details: We adopt the two-stage top-down human pose estimation pipeline. Firstly, the person instances are detected, and then the keypoints are estimated. In the stage of keypoint estimation, we set the cropped human images to low resolution. We employ a commonly used person detector provided by SimpleBaselines [33] with 56.4% AP. Simcc is utilized as the base model for the teacher and the student, which uses HRNet [9] as its backbone.

We use the Adam optimizer [80], the base learning rate is initialized as $1e$-3. It is subsequently reduced to $1e$-4 and $1e$-5 at the 170-th and 200-th epochs respectively. The data augmentation for CDKD is the same as Simcc. Experiments are conducted on an NVIDIA GeForce RTX 2080 Ti GPU and an NVIDIA GeForce RTX 3080 Ti GPU.

All methods, including our proposed approach and the SOTA methods used for comparison, are trained and tested under low-resolution conditions. The resolution used for train-

TABLE II: Comparison with SOTA methods on the COCO validation dataset. (*) denotes the experimental results obtained by incrementing the splitting factor's value followed by Simcc [21].

Method	Input size	Backbone	AP↑	AP^{50}↑	AP^{75}↑	AP^{M}↑	AP^{L}↑	AR↑	#Params↓	GFLOPs↓
HRNet [9]	32×32	HRNet-W32	8.2	36.9	1.3	9.2	6.8	15.0	28.5M	0.1
TokenPose [11]		HRNet-W32	14.0	48.2	3.4	15.2	12.5	21.7	33.1M	0.2
CAL [22]		HRNet-W32	26.4	61.9	18.2	27.1	25.7	33.9	49.3M	0.2
Rle [25]		HRNet-W32	24.4	-	-	-	-	-	29.1M	0.6
Dark [36]		HRNet-W32	12.5	45.2	2.5	13.8	11.1	20.3	28.5M	0.1
SimBase [33]		ResNet-152	4.4	21.1	1.0	5.3	3.2	9.0	68.6M	0.3
PCT [1]		Swin-Base	1.3	4.6	10.0	1.0	1.2	3.1	86.9M	1.7
Distillpose [2]		HRNet-W32	9.5	32.6	2.4	10.0	9.5	21.2	33.0M	0.3
SimCC (*baseline*) [21]		HRNet-W32	29.8	65.6	22.5	30.0	29.9	36.3	28.5M	0.1
CDKD (*ours*)		HRNet-W32	**30.7**$_{+0.9}$	**66.4**	**23.5**	**30.9**	**30.7**	**37.3**	28.5M	0.1
SimBase [33]	64×64	ResNet-152	30.3	67.6	22.6	30.6	30.5	36.2	68.6M	1.3
Distillpose [2]		HRNet-W32	31.7	66.8	26.6	32.3	31.8	44.5	33.1M	0.8
PCT [1]		Swin-Base	13.8	41.1	6.4	14.1	14.2	19.6	86.9M	2.3
Rle [25]		HRNet-W32	52.5	-	-	-	-	-	29.1M	0.6
HRNet [9]		HRNet-W48	46.9	83.7	49.2	46.6	47.5	52.6	63.6M	1.2
Tokenpose [11]		HRNet-W48	50.3	82.7	54.4	49.9	51.4	55.6	68.2M	1.2
CAL [22]		HRNet-W48	60.6	**88.1**	68.4	59.5	62.3	65.5	110.3M	1.8
Dark [36]		HRNet-W48	57.2	86.8	63.5	55.9	59.2	62.2	63.6M	1.2
Simcc (*baseline*) [21]		HRNet-W48	58.6	85.9	64.9	57.8	60.5	63.4	63.7M	1.2
CDKD (*ours*)		HRNet-W48	60.3$_{+1.7}$	86.9	67.3	59.6	62.0	65.0	63.7M	1.2
Simcc (*baseline*)* [21]		HRNet-W48	59.7	85.0	67.3	58.4	**64.0**	**67.5**	63.7M	1.2
CDKD (*ours*)*		HRNet-W48	**61.1**$_{+1.4}$	86.9	**68.4**	**59.9**	62.9	65.6	63.7M	1.2

TABLE III: Comparisons with the baseline method on the CrowdPose test dataset.

Scheme	Backbone	Input size	AP↑	AP^{50}↑	AP^{75}↑	AP^{E}↑	AP^{M}↑	AP^{H}↑	#Params↓	GFLOPs↓
Simcc (*baseline*) [21]	HRNet-W32	64×64	46.5	70.9	50.0	51.2	43.1	31.8	63.7M	1.2
CDKD (*ours*)	-	64×64	**47.0**	**72.3**	**50.6**	**57.7**	**48.0**	**34.4**	63.7M	1.2
Simcc (*baseline*) [21]	-	32×32	23.3	52.5	17.6	31.4	23.8	14.7	28.8M	0.1
CDKD (*ours*)	-	32×32	**23.7**	**52.8**	**18.1**	**31.4**	**24.3**	**15.4**	28.8M	0.1

ing and testing is the same. For the selected SOTA methods, we only modify the input resolution, while keeping all other experimental settings consistent with those in the original papers. They do not employ any additional distillation. In our method, we first train a teacher model under high-resolution conditions. Then, we train a student model under low-resolution conditions, distilling knowledge from the high-resolution teacher to the low-resolution student. In this paper, all teacher models are trained with an input resolution of 256×256.

B. Main Results

1) Results on MPII Dataset.: We evaluate the CDKD framework on the MPII validation dataset. Table I compares the PCKh@0.5 accuracy results of CDKD and the SOTA methods under low-res conditions. We can clearly observe that Simcc achieves outstanding performance with low computational cost. Adding CDKD, Simcc is further improved. It achieves better accuracy than any other method. Specifically, in resolutions of 64×64 and 32×32, our method improves the baseline by 1.1% and 1.8%, respectively. It does not incur any additional costs in parameters and GFLOPs, The performance is significant as compared to prior works. In short, CDKD achieves SOTA performance with ***NO*** increase in computational cost.

2) Results on COCO Dataset.: Table II shows the results of the cutting-edge methods and CDKD under low-res conditions on the COCO validation set. In resolutions of 32×32 and 64×64, our method improves the baseline model 0.9% and 1.7%, respectively. Especially, it achieves SOTA performance, whether in a resolution of 32×32 or 64×64. Similarly, it incurs no additional computational cost or parameter increase, providing additional evidence for the efficiency and effectiveness of our proposed CDKD method for low-res HPE.

3) Results on Crowdpose Dataset.: We further demonstrate the effectiveness of the proposed CDKD method on the CrowdPose dataset. Following the original work [24], we use YoloV3 [81] as the human detector, with a batch size set to 256. Comparative experiments were conducted on the CrowdPose test set with input sizes of 64×64 and 32×32, respectively. The results in Table III show that the CDKD-based method outperforms the baseline method. It demonstrates that CDKD remains effective on the CrowdPose dataset.

TABLE IV: Ablation studies for different modules on the MPII (64×64) and COCO (32×32) validation sets. The "NECM" means the non-ETHT CCA module. All experiments use HRNet-W32 as the backbone. On the MPII dataset, we independently train and evaluate each method five times, and report the results as "mean±standard deviation".

Method	SAPE			CCA		AP↑	PCKh@0.5↑
	Projector	Ensemble	SAU	NECM	ETHT		
Simcc *(baseline)*	-	-	-	-	-	29.8	78.37±0.02
ours	✓					30.3	78.73±0.10
ours	✓	✓				30.4	78.88±0.12
ours	✓	✓	✓			30.6	78.95±0.15
ours				✓		30.3	78.68±0.24
ours				✓	✓	30.6	78.85±0.18
ours	✓	✓	✓	✓	✓	**30.7**	**79.24±0.21**

TABLE V: Ablation studies of the alignment method on the MPII validation dataset. All experiments use HRNet-W32 as the backbone. "Conv Downsample" and "Fully Connected Layer" mean conventional feature and class alignment methods, respectively. "Same Class Number" refers to the method of making the number of categories consistent between the teacher model and the student model by altering the splitting factor [21].

Distillation	Method	PCKh@0.5↑
Feature distillation	Conv Downsample [71]	78.2
	Fully Connected Layer [52]	78.2
	SAPE *(ours)*	**78.8**
Logit distillation	Same Class Number [21]	75.2
	Fully Connected Layer [82]	78.0
	CCA *(ours)*	**78.8**

In summary, our proposed CDKD consistently achieves significant improvements across different datasets, which provides strong evidence for the effectiveness of CDKD.

C. Ablation Study

Different modules. In this subsection, we conduct several ablation experiments to show how each module helps the training of the low-res student. As shown in Table IV, on the COCO val dataset, all modules benefit the low-res model. SAPE and CCA bring an improvement of 0.8% and 0.8%, respectively. In SAPE, the projector ensemble yields a 0.6% performance gain, and the scale-adaptive unit (SAU) further improves the performance of the projector ensemble by 0.2%. Meanwhile, the NECM brings the student 0.5% gains. When combing the ETHT strategy, the gains get to 0.8%. The combination of all proposed modules bring the best performance, which improves the performance by 0.9%.

In order to thoroughly demonstrate the performance of the various modules, we also conduct ablation experiments on the MPII dataset. To reduce the impact of randomness, we conduct 5 independent repeated experiments, and report the mean and standard deviation of the results in Table IV. The average performance consistently improves with the inclusion of each module. Together with the results on the COCO dataset, this provides strong evidence that each module makes a valuable contribution to the model's overall performance. The combination of all modules achieves the optimal performance for the model.

TABLE VI: Ablation studies of the number of convolutions with different kernel sizes in SAU on the MPII validation dataset. HRNet-W32 is used as the backbone. The 3×3, 5×5, 7×7, and 9×9 refer to the conv kernel sizes.

Conv	3×3	5×5	7×7	9×9	PCKh@0.5↑
Conv number	1	1	1	0	79.0
	2	1	1	0	78.9
	3	1	1	0	78.5
Conv number	1	2	1	0	79.2
	1	1	2	0	**79.4**
	1	1	1	1	79.0

TABLE VII: Ablation studies on the hyperparameters of the ETHT strategy are conducted on the MPII validation dataset using the HRNet-W32 backbone. The τ, α, and β refer to key hyperparameters in the CDKD framework.

Method	Hyperparameters	PCKh@0.5↑
Non-ETHT	-	78.8
ETHT	τ	**79.4**
ETHT	$\tau,\ \alpha$	79.1
ETHT	$\tau,\ \alpha,\ \beta$	79.2

Alignment methods. We compare the results of CDKD with conventional feature alignment and class alignment in Table V. As shown in the table, SAPE outperforms the fully-connected layer by 0.6%, while CCA outperforms it by 0.8%. It indicates our proposed approach surpasses traditional alignment methods.

Conv number. We compare the impact of the convolution number of different kernel sizes in our proposed SAU module, as shown in Table VI. we can observe that increasing moderately the proportion of convolutions with a large receptive field can further improve the performance of our proposed CDKD distillation framework. It demonstrates that our designed module adequately captures multi-scale human body information.

Hyperparameters. Table VII shows how the performance of our proposed framework is affected by the choice of hyperparameters in Eq. 7 and Eq. 8. It indicates that the ETHT strategy we designed is effective. Each choice of hyperparameters helps to improve the model performance. When selecting τ as a learning object, our method reaches the best result.

Projector number. We evaluate the impact of projector number on the performance of CDKD in Fig. 6. The use of a projector ensemble improves the model's performance, but an excessive number of projectors would result in a decline in performance with the adoption of our SAU. We conducted experiments with a deeper SAU module to further investigate the causes of performance degradation associated with increas-

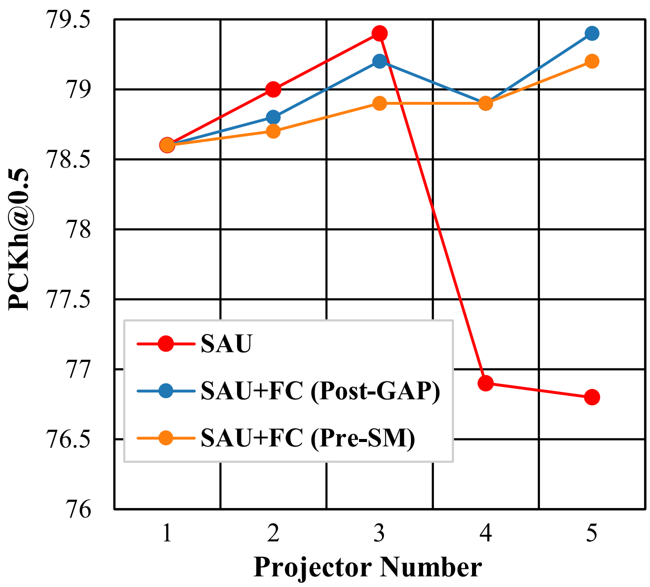

Fig. 6: Ablation study on the number of projectors on the MPII validation set. The projector number denotes the number of projectors used in the SAPE module. “SAU+FC (Post-GAP)” refers to the addition of a fully connected layer after the global average pooling (GAP) layer, while “SAU+FC (Pre-SM)” refers to the addition of a fully connected layer before the softmax layer.

TABLE VIII: Evaluation of the CCA module within our proposed CDKD framework across different models on the COCO validation dataset.

Decoding Method	Resolution	Role	Backbone	Class Number	AP↑
Classification	256×256	teacher	HRNet-48	768	76.4
-	64×64	baseline	ResNet-50	192	40.0
-	64×64	student	ResNet-50	192	**42.0**$_{+2.0}$

ing the number of projectors. As illustrated in Fig. 6, fully connected layers are introduced at various positions to conduct an ablation study on the number of projectors. The results indicate that, within a deeper SAU, performance improves as the number of projectors increases. However, this improvement is not significant and, conversely, leads to increased computational overhead. Notably, employing three projectors in conjunction with our proposed SAU module achieves the optimal performance while incurring lower training costs.

D. Universality

We further extend our CDKD framework to different pose estimation models to illustrate its universality. In current HPE research, there is a clear difference in the applicability of the CCA and SAPE modules. The CCA module is only suitable for models that include classification operations, whereas SAPE can be applied to almost all models.

Therefore, we conduct separate research on the CCA module and the SAPE module. We first reproduced both the high-resolution teacher model and the low-resolution student model. Next, we trained the low-resolution student model incorporating our proposed method.

TABLE IX: Evaluation of the proposed SAPE module applied to different models on the MPII validation set. The teacher models for CAL and HRNet are both trained at a resolution of 256×256.

Method	Backbone	Resolution	PCKh@0.5↑
CAL	HRNet-W32	32×32	55.4
CAL (+SAPE)	-	32×32	**56.0**$_{+0.6}$
CAL (Teacher)	-	256×256	90.4
HRNet	-	64×64	74.8
HRNet (+SAPE)	-	64×64	**76.2**$_{+1.4}$
HRNet (Teacher)	-	256×256	90.3

CCA module. We extend the CCA module to different models. As shown in Table VIII, the teacher model and the student model employ distinct backbones and have varying numbers of classes. The two models differ in all aspects except for the decoding method. In this case, the student model is still well-optimized by the CCA module, demonstrating the universality of our CCA module.

SAPE module. As an easy-to-use plug-in technique, SAPE can be seamlessly integrated into existing HPE works. For different HPE models, it is feasible to distill feature knowledge from teacher models into low-resolution student models. First, we train a high-resolution teacher model. Then, we distill knowledge from the teacher to the student model via the SAPE module during the student model training. We only add the SAPE-based distillation, while all other settings remain consistent with the original configuration. Except for the input resolution, the teacher and student models share the same architecture and settings. As shown in Table IX, our method achieves improvements on the two cutting-edge methods, fully demonstrating its generality and effectiveness.

E. Qualitative Analysis

1) Visualization of Pose Estimation Results: Fig. 7 provides visualized HPE results on the COCO validation dataset. As is shown in Fig. 7, the CDKD model achieves accurate human pose estimation when the input is low-resolution.

To further demonstrate the robustness of CDKD, we present pose estimation results in real-world scenarios, as shown in Fig. 8. The predictions are made using a model trained on the COCO train dataset, and the test images are selected from the MPII, CrowdPose, and ADE20K datasets, which better reflect in-the-wild conditions. Training and testing on different datasets provide stronger evidence of the model's generalization ability. We adopt HRNet-W48 as the backbone of our model, which is trained on input images resized to 64×64. As shown in Fig. 8, CDKD maintains stable and accurate prediction performance across various challenging scenarios.

2) Visualization of Attention Maps: To intuitively demonstrate the effectiveness of CDKD, we visualize and compare the convolutional attention maps of both CDKD and the baseline. These samples are selected from the MPII validation

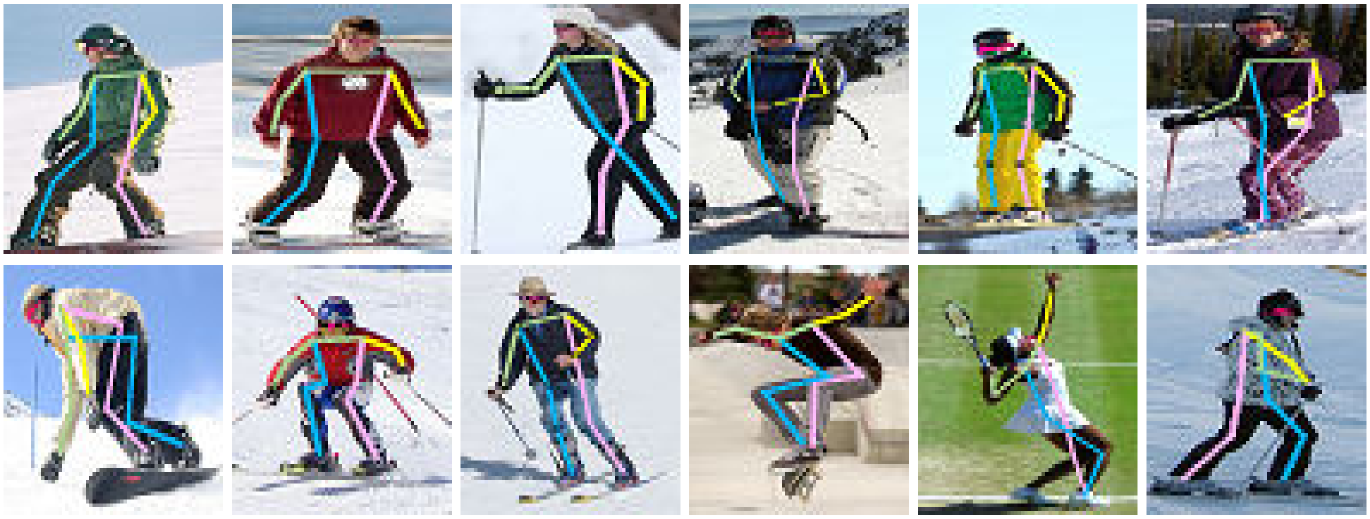

Fig. 7: Visual results of our CDKD framework on the COCO validation dataset. Our base model is SimCC with an HRNet-W48 backbone. The resolution of all images is 64×64.

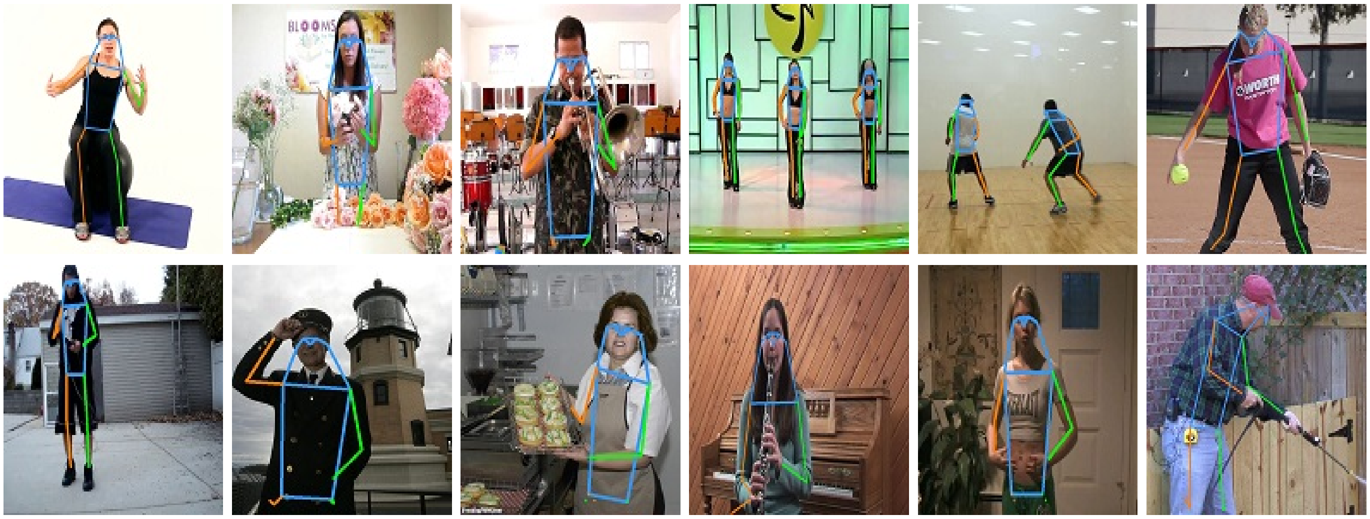

Fig. 8: Qualitative results on in-the-wild low-resolution scenarios.

TABLE X: Comparisons of the training complexity on the MPII train set. The input resolution is 64×64, and the backbone is HRNet-W32. All experiments are conducted on a single NVIDIA 2080 Ti GPU.

Method	#Params	GFLOPs	Time per Epoch	GPU Memory	Converged Epoch
Simcc *(baseline)*	28.6M	0.6	142s	2224MB	168
CDKD	63.5M	10.1	180s	8904MB	170

set, and experiments are conducted on different keypoints and various convolutional layers. As shown in Fig. 9, CDKD consistently exhibits a stronger focus on target keypoint regions compared to the baseline. This indicates that the high-resolution knowledge provided by the teacher model effectively guides the student model to attend to critical regions. These visualizations not only validate the effectiveness of CDKD but also enhance the interpretability of the method.

F. Training Complexity Analysis

We conduct several experiments to analyze the training complexity of CDKD. As shown in Table X, CDKD exhibits higher training costs compared to the baseline in terms of parameter count, computational complexity, memory consumption, and training time. However, knowledge distillation increases model complexity solely during training, incurring no extra computational overhead at inference time. Under the same deployment

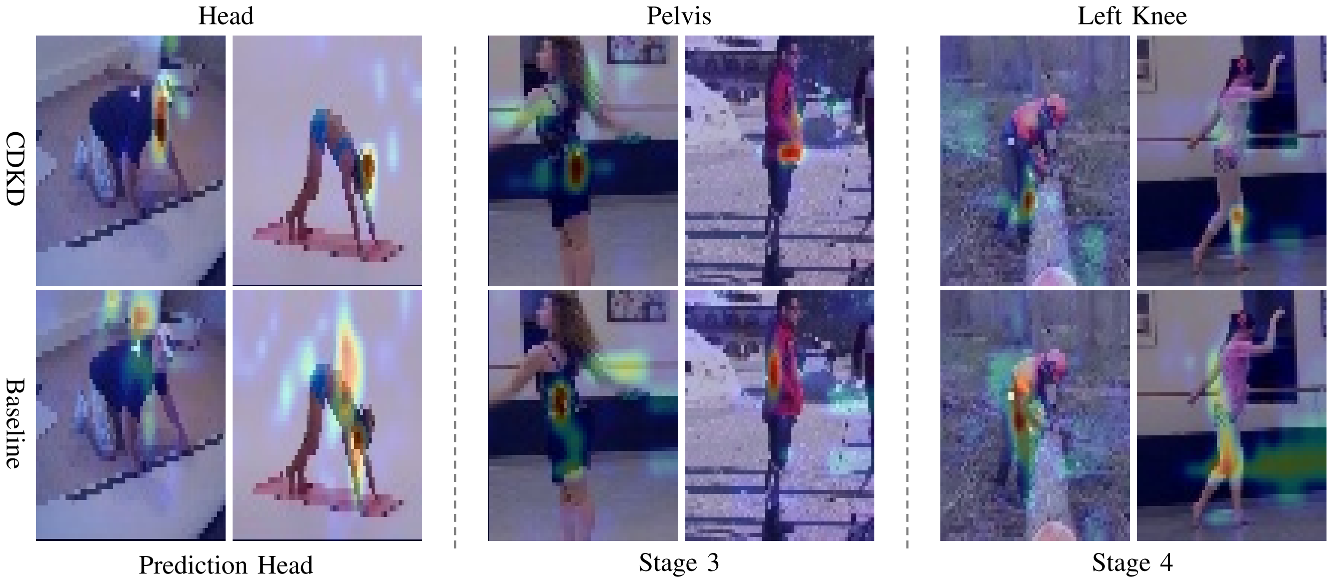

Fig. 9: Visual comparison of convolutional attention maps between CDKD and the baseline. “Prediction Head”, “Stage 3”, and “Stage 4” correspond to different stages of HRNet. We choose the final convolutional layer from each of these stages for visualization. “Head”, “Pelvis”, and “Left Knee” indicate the target keypoints attended by the model. The images are from the MPII validation dataset with a resolution of 64×64.

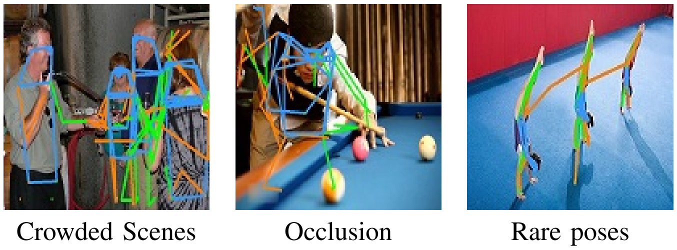

Fig. 10: Visualization of failure cases.

cost, CDKD achieves superior performance over the baseline. A relatively expensive single training phase brings permanent performance improvements without increasing inference cost, making it a worthwhile optimization strategy.

G. Failure Case Analysis

Figure 10 illustrates the failure cases of our CDKD, highlighting its limitations in handling challenging scenarios such as occlusion, crowding, and complex poses. Under low-resolution conditions, these scenarios are often difficult to handle. Occlusions obscure local keypoint information; crowded scenes introduce interference between subjects, hampering accurate localization; and the scarcity of complex poses in the training data limits the model’s ability to generalize to such cases. These factors collectively degrade the accuracy of low-resolution models. In future work, we plan to address these issues by leveraging richer human structural priors and increasing the number of complex pose samples.

V. Limitation and Future Work

As a novel universal distillation framework, the proposed CDKD opens up many possible directions for future works to address the bottleneck limitations. Our SAPE module can be extended to various knowledge distillation tasks, enabling the distillation between features with different sizes. The CCA module we propose addresses the issue of inconsistent categories. It opens up a new perspective on distillation across categories. Nonetheless, our cross-domain knowledge distillation method is presently confined to distilling between conventional models. In the future, we plan to apply our method to large models, thus maximizing the utilization of their abundant knowledge.

VI. Conclusion

In this work, we propose a novel low-resolution human pose estimation framework (CDKD), which includes a scale-adaptive projector ensemble (SAPE) module and a cross-class alignment (CCA) module to perform high-resolution to low-resolution knowledge distillation. In this way, the student model acquires richer image knowledge at both feature and logit levels, achieving a big leap in performance while maintaining efficiency. Extensive experiments conducted on the COCO, MPII, and Crowdpose datasets demonstrate the effectiveness of CDKD. In short, our proposed CDKD achieves SOTA performance among low-res methods with a low computational cost.

References

[1] Z. Geng, C. Wang, Y. Wei, Z. Liu, H. Li, and H. Hu, “Human pose as compositional tokens,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 660–671.

[2] S. Ye, Y. Zhang, J. Hu, L. Cao, S. Zhang, L. Shen, J. Wang, S. Ding, and R. Ji, “Distilpose: Tokenized pose regression with heatmap distillation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2163–2172.

[3] H. Qu, L. Xu, Y. Cai, L. G. Foo, and J. Liu, "Heatmap distribution matching for human pose estimation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 327–24 339, 2022.

[4] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 571–38 584, 2022.

[5] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Keypoint localization via transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 802–11 812.

[6] W. Yang, W. Ouyang, H. Li, and X. Wang, "End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3073–3082.

[7] U. Iqbal, P. Molchanov, and J. Kautz, "Weakly-supervised 3d human pose learning via multi-view images in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5243–5252.

[8] L. Schmidtke, A. Vlontzos, S. Ellershaw, A. Lukens, T. Arichi, and B. Kainz, "Unsupervised human pose estimation through transforming shape templates," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2484–2494.

[9] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[10] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[11] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.-T. Xia, and E. Zhou, "Tokenpose: Learning keypoint tokens for human pose estimation," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 11 313–11 322.

[12] Y. Cheng, B. Wang, and R. T. Tan, "Dual networks based 3d multi-person pose estimation from monocular video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1636–1651, 2022.

[13] H. Liu, T. Liu, Y. Chen, Z. Zhang, and Y.-F. Li, "Ehpe: Skeleton cues-based gaussian coordinate encoding for efficient human pose estimation," *IEEE Transactions on Multimedia*, 2022.

[14] W. Li, H. Liu, R. Ding, M. Liu, P. Wang, and W. Yang, "Exploiting temporal contexts with strided transformer for 3d human pose estimation," *IEEE Transactions on Multimedia*, vol. 25, pp. 1282–1293, 2022.

[15] M. Li, Z. Zhou, and X. Liu, "Multi-person pose estimation using bounding box constraint and lstm," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2653–2663, 2019.

[16] S. Zou, X. Zuo, S. Wang, Y. Qian, C. Guo, and L. Cheng, "Human pose and shape estimation from single polarization images," *IEEE Transactions on Multimedia*, 2022.

[17] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1246–1259, 2017.

[18] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[19] Y. Chen, S. Wang, J. Liu, X. Xu, F. de Hoog, and Z. Huang, "Improved feature distillation via projector ensemble," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 084–12 095, 2022.

[20] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: many could be better than all," *Artificial intelligence*, vol. 137, no. 1-2, pp. 239–263, 2002.

[21] Y. Li, S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, W. Yang, and S.-T. Xia, "Simcc: A simple coordinate classification perspective for human pose estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 89–106.

[22] C. Wang, F. Zhang, X. Zhu, and S. S. Ge, "Low-resolution human pose estimation," *Pattern Recognition*, vol. 126, p. 108579, 2022.

[23] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[24] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 863–10 872.

[25] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, and C. Lu, "Human pose regression with residual log-likelihood estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 025–11 034.

[26] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, and Z. Tu, "Pose recognition with cascade transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1944–1953.

[27] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2830–2838.

[28] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4733–4742.

[29] Z. Tian, H. Chen, and C. Shen, "Directpose: Direct end-to-end multi-person pose estimation," *arXiv preprint arXiv:1911.07451*, 2019.

[30] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653–1660.

[31] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," *arXiv preprint arXiv:1901.00148*, 2019.

[32] Y. Wang, M. Li, H. Cai, W.-M. Chen, and S. Han, "Lite pose: Efficient architecture design for 2d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 126–13 136.

[33] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481.

[34] A. Bulat and G. 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*. Springer, 2016, pp. 717–732.

[35] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *Advances in neural information processing systems*, vol. 27, 2014.

[36] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7093–7102.

[37] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[38] X. Nie, J. Feng, J. Zhang, and S. Yan, "Single-stage multi-person pose machines," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6951–6960.

[39] A. Varamesh and T. Tuytelaars, "Mixture dense regression for object detection and human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 086–13 095.

[40] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5700–5709.

[41] H. Wang, J. Liu, J. Tang, and G. Wu, "Lightweight super-resolution head for human pose estimation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2353–2361.

[42] J. Yang, A. Zeng, S. Liu, F. Li, R. Zhang, and L. Zhang, "Explicit box detection unifies end-to-end multi-person pose estimation," *arXiv preprint arXiv:2302.01593*, 2023.

[43] S. Lee, J. Rim, B. Jeong, G. Kim, B. Woo, H. Lee, S. Cho, and S. Kwak, "Human pose estimation in extremely low-light conditions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 704–714.

[44] X. Ju, A. Zeng, J. Wang, Q. Xu, and L. Zhang, "Human-art: A versatile human-centric dataset bridging natural and artificial scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 618–629.

[45] Z. Yang, A. Zeng, C. Yuan, and Y. Li, "Effective whole-body pose estimation with two-stages distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4210–4220.

[46] Q. Sun, Y. Wang, A. Zeng, W. Yin, C. Wei, W. Wang, H. Mei, C. S. Leung, Z. Liu, L. Yang *et al.*, "Aios: All-in-one-stage expressive human pose and shape estimation," *arXiv preprint arXiv:2403.17934*, 2024.

[47] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang *et al.*, "Smpler-x: Scaling up expressive

human pose and shape estimation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 11 454–11 468, 2023.

[48] A. Kumar and R. Chellappa, "S2ld: Semi-supervised landmark detection in low-resolution images and impact on face verification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 758–759.

[49] J. C. L. Chai, T.-S. Ng, C.-Y. Low, J. Park, and A. B. J. Teoh, "Recognizability embedding enhancement for very low-resolution face recognition and quality estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9957–9967.

[50] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022, pp. 443–459.

[51] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, and F. De la Torre, "3d human shape and pose from a single low-resolution image with self-supervised learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 284–300.

[52] Y. Zhu, Q. Zhou, N. Liu, Z. Xu, Z. Ou, X. Mou, and J. Tang, "Scalekd: Distilling scale-aware knowledge in small object detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 723–19 733.

[53] Z. Ni, F. Yang, S. Wen, and G. Zhang, "Dual relation knowledge distillation for object detection," *arXiv preprint arXiv:2302.05637*, 2023.

[54] W. Huang, Z. Peng, L. Dong, F. Wei, J. Jiao, and Q. Ye, "Generic-to-specific distillation of masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 996–16 005.

[55] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962.

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

[57] J. Guo, M. Chen, Y. Hu, C. Zhu, X. He, and D. Cai, "Reducing the teacher-student gap via spherical knowledge disitllation," *arXiv preprint arXiv:2010.07485*, 2020.

[58] Y. Niu, L. Chen, C. Zhou, and H. Zhang, "Respecting transfer gap in knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 21 933–21 947, 2022.

[59] Y. Jin, J. Wang, and D. Lin, "Multi-level logit distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 276–24 285.

[60] Z. Zheng, R. Ye, P. Wang, D. Ren, W. Zuo, Q. Hou, and M.-M. Cheng, "Localization distillation for dense object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9407–9416.

[61] R. Miles, I. Elezi, and J. Deng, "v_kd : improving knowledge distillation using orthogonal projections," *arXiv preprint arXiv:2403.06213*, 2024.

[62] J. Wang, Y. Chen, Z. Zheng, X. Li, M.-M. Cheng, and Q. Hou, "Crosskd: Cross-head knowledge distillation for object detection."

[63] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1921–1930.

[64] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3779–3787.

[65] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[66] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[67] L. Qi, J. Kuen, J. Gu, Z. Lin, Y. Wang, Y. Chen, Y. Li, and J. Jia, "Multi-scale aligned distillation for low-resolution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 443–14 453.

[68] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[69] S. Shin, J. Lee, J. Lee, Y. Yu, and K. Lee, "Teaching where to look: Attention similarity knowledge distillation for low resolution face recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 631–647.

[70] Z. Huang, S. Yang, M. Zhou, Z. Li, Z. Gong, and Y. Chen, "Feature map distillation of thin nets for low-resolution object recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 1364–1379, 2022.

[71] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[72] X. Deng and Z. Zhang, "Comprehensive knowledge distillation with causal intervention," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 158–22 170, 2021.

[73] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 510–519.

[74] R. Diaz and A. Marathe, "Soft labels for ordinal regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4738–4747.

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[76] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.

[77] Z. Li, J. Ye, M. Song, Y. Huang, and Z. Pan, "Online knowledge distillation for efficient pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 11 740–11 750.

[78] Z. Li, X. Li, L. Yang, B. Zhao, R. Song, L. Luo, J. Li, and J. Yang, "Curriculum temperature for knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1504–1512.

[79] F. Zhang, X. Zhu, and M. Ye, "Fast human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3517–3526.

[80] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[81] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[82] S. Ge, S. Zhao, C. Li, Y. Zhang, and J. Li, "Efficient low-resolution face recognition via bridge distillation," *IEEE Transactions on Image Processing*, vol. 29, pp. 6898–6908, 2020.