Regularizing Vector Embedding in Bottom-Up Human Pose Estimation

Haixin Wang[1,2] (✉), Lu Zhou[2], Yingying Chen[2], Ming Tang[1,2], and Jinqiao Wang[1,2,3,4]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[3] Peng Cheng Laboratory
[4] ObjectEye Inc.
{haixin.wang, lu.zhou, yingying.chen, tangm, jqwang}@nlpr.ia.ac.cn

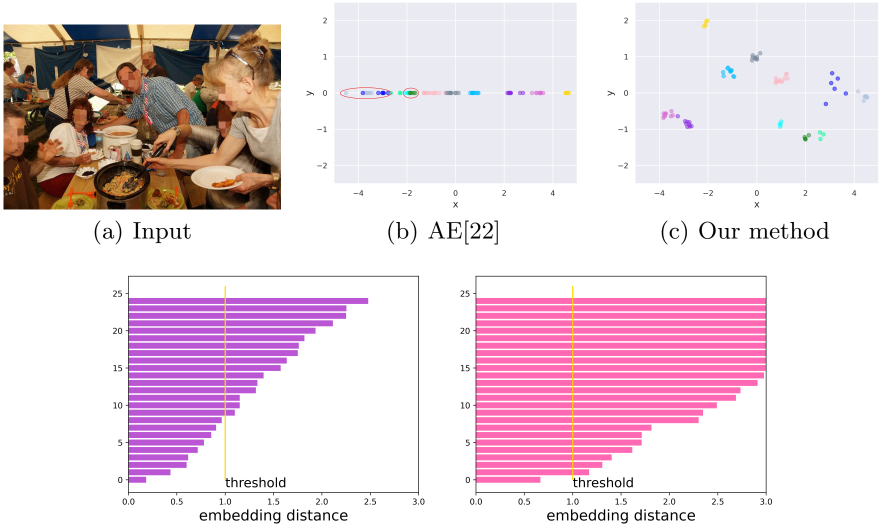
Abstract. The embedding-based method such as Associative Embedding is popular in bottom-up human pose estimation. Methods under this framework group candidate keypoints according to the predicted identity embeddings. However, the identity embeddings of different instances are likely to be linearly inseparable in some complex scenes, such as crowded scene or when the number of instances in the image is large. To reduce the impact of this phenomenon on keypoint grouping, we try to learn a sparse multidimensional embedding for each keypoint. We observe that the different dimensions of embeddings are highly linearly correlated. To address this issue, we impose an additional constraint on the embeddings during training phase. Based on the fact that the scales of instances usually have significant variations, we uilize the scales of instances to regularize the embeddings, which effectively reduces the linear correlation of embeddings and makes embeddings being sparse. We evaluate our model on CrowdPose Test and COCO Test-dev. Compared to vanilla Associative Embedding, our method has an impressive superiority in keypoint grouping, especially in crowded scenes with a large number of instances. Furthermore, our method achieves state-of-the-art results on CrowdPose Test (74.5 AP) and COCO Test-dev (72.8 AP), outperforming other bottom-up methods. Our code and pretrained models are available at https://github.com/CR320/CoupledEmbedding.

Keywords: human pose estimation, bottom-up, embedding

1 Introduction

Multi-person human pose estimation (HPE) is a fundamental task in computer vision. Current multi-person HPE methods are mainly split up into two paradigms: top-down and bottom-up. Top-down methods [25,30,3,28] first detect instances via human detector and then perform keypoint detection for each detected instance. By contrast, bottom-up methods [2,22,24,14] first detect all identity-free keypoints and then group them into individual persons.

Judging the identities of candidate keypoint is a significant challenge of bottom-up methods. The part field-based methods [2,14] utilize limb information to construct connective intensity between keypoints. The human center regression-based methods [23,27,8] utilize a human center point to represent the instance and densely estimate keypoint offsets w.r.t. the center. The embedding-based methods [22,4,21] assign each candidate keypoint an identity embedding and group keypoints with a heuristic matching algorithm in post-processing. In recent years, the embedding-based methods are popular in human pose estimation.



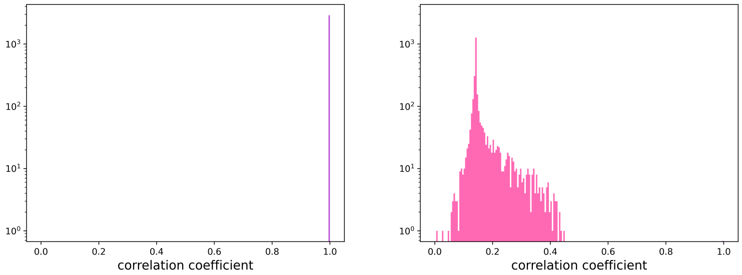
(a) Input (b) AE[22] (c) Our method

(d) AE[22] (e) Our method

Fig. 1. (a) shows the input image in a typical complex scene. (b) and (c) show the embedding distribution in a 2-dimensional space (the original embedding dimensionality is set to 8, we adopt PCA to reduce the dimensionality to 2). (d) and (e) show the embedding distances between different instances (we only show the smallest 25 pairs)

Despite the embedding-based methods achieve impressive results on some common benchmarks [1,20], they still suffer from precision degradation in some complex scenes. For example, in crowded scenes with a large number of instances, some predicted embeddings of different instances are likely to be linearly inseparable, resulting in incorrect keypoint grouping in the post-processing of the embedding-based methods. When the 1-D embeddings are replaced by multidimensional embeddings, they are still distributed on a line and the linear inseparability is not diminished. As shown in Figure 1(b), the embddings predicted by AE [22] in red circle are linearly inseparable. To analyze the relationship between different dimensions of embeddings, we collect images with more than

3 instances in CrowdPose Test [18] and calculate the correlation coefficients between different dimensions of embeddings in one image. The histogram in Figure 2(a) shows that all the mean correlation coefficients are really close to 1, which means different dimensions of embeddings predicted by AE [22] are highly linearly correlated.



(a) AE[22]

(b) Our method

Fig. 2. Histogram of correlation coefficients. (a) shows the result of AE [22] and (b) shows the result of our method. The dimensionalities of embeddings are set to 8 in both methods. All samples are collected from CrowdPose Test [18].

To learn sparse multidimensional embeddings, we propose a novel method named Coupled Embedding (CE). Different from vanilla embedding-based methods, our method imposes an additional regularization on the embeddings. Based on the fact that the scales of instances usually have significant variations, we choose the scales as the supervised inforamtion for regularization. In our method, we convert the instance scales to a normalized vector which only has high values in one or two adjacent dimensions. Then we minimize the angle between the scale vector and the vector embedding. This constraint pushes the principal components of embeddings to be concentrated in certain dimensions. Figure 2(b) shows that our method significantly reduces the linear correlation of embeddings in an images.

To demonstrate the superiority of our method, we compare the embeddings of AE [22] with our method in a typical complex scene. Comparing the result in Figure 1(a) and Figure 1(b), it shows embeddings in AE [22] are almost distributed on a line but the embeddings in our methods are scattered. Some linearly inseparable embeddings predicted by AE [22] become linearly separable in our method. Figure 1(d) shows that some embedding distances between different instances in AE [22] are lower than grouping threshold, which will cause incorrect keypoint grouping. By contrast, Figure 1(e) shows that the embedding distances increase apparently in our method, which means our method can efficiently improve keypoint grouping.

Recent SOTA embedding-based methods [4,21] focus on improving heatmap regression to enhance keypoint detection. As keypoint grouping is based on the results of keypoint detection, for a fair comparison with current embedding-based methods, we further utilize learned scale information and adaptive loss weights to improve heatmap regression.

In conclusion, our contributions are mainly as follows:

- To our best knowledge, we are the first to probe the limitation of keypoint grouping in Associative Embedding. We find that different dimensions of embeddings in an image are highly linearly correlated.
- We propose a novel keypoint grouping method named Coupled Embedding for bottom-up human pose estimation. Our method imposes an additional constraint on embeddings to learn sparse multidimensional embeddings.
- Our method achieves new state-of-the-art results on CrowdPose Test (74.5 AP) and COCO Test-dev (72.8 AP), outperforming all existing bottom-up methods. We conduct a series of experiments and the results demonstrate that our model has significant advantages in crowded scenes compared with other bottom-up methods.

2 Related Work

2.1 Bottom-Up Methods

Bottom-up methods first detect all candidate keypoints in an image, then assemble them into each instance. Pioneering works such as DeepCut [26] and L-JPA [12] formulate the keypoint association problem as an integer linear program, which however takes longer processing time. In recent years, there are three popular types of bottom-up methods, including part field [2,14,17], human center regression [23,27,8] and identity embedding [22,4,21]. The part field-based methods produce a 2D vector field to construct connective intensity between keypoints. OpenPose[2] is a representative part field-based method that predicts the part affinity fields to construct the connective intensity and then utilizes a greedy algorithm to assemble different keypoints of the same instance. Inspired by OpenPose [2], PifPaf [14] utilizes the part intensity fields to localize body parts, and employs the part association fields to associate body parts with each other. The human center regression-based methods first locate a center position of each instance, then densely predict displacements w.r.t the center position for keypoints which belong to the instance. The method [23] proposes a single-stage multi-person pose machine that simultaneously regresses the center positions and body keypoint displacements, predicting multi-person poses within one stage. DEKER [8] utilizes a multi-branch structure for separate regression, where each branch learns a representation with dedicated adaptive convolutions and regresses one keypoint offset. Associative Embedding [22] is the first to predict identity embeddings for keypoint grouping. Later methods [4,21] focus on improving heatmap regression to enhance keypoint detection. Higherhrnet [4] utilize higher-resolution heatmaps to handle scale variation. The method [21] add a new branch to predict the uncertainty maps which adaptively adjust the standard deviation of the gaussian kernel for each keypoint, enabling the model to be more tolerant of various human scales and labeling ambiguities.

2.2 Vector Embedding

Prior works apply vector embedding to many tasks. The methods [6,29] in image retrieval utilize vector embedding to measure similarity between images. The methods [31,9] map visual features and text features to the same vector space to establish their connection in image classification or image captioning. Deep clustering method [7,11] utilize vector embedding to obtain a feasible feature space. Recently, many methods [22,4,21] in human pose estimation and object detection [16] apply vector embedding in keypoint grouping.

For multi-person human pose estimation, Associative Embedding [22] is the first to apply vector embedding for keypoint grouping and defines a loss function which prompts keypoints from the same instance to have similar embeddings and keypoints across different instances to have distinguishing embeddings. The authors of Associative Embedding [22] found there are little performance gap between 1-D embedding and multi-dimensional embedding, but they did not probe the reason further. In our paper, we find that the different dimensions of embeddings in an image are highly linearly correlated. Therefore, even in high-dimensional space, the embeddings are almost distributed on one line, which is the same as the 1-D case.

3 Our Method

3.1 Model Framework

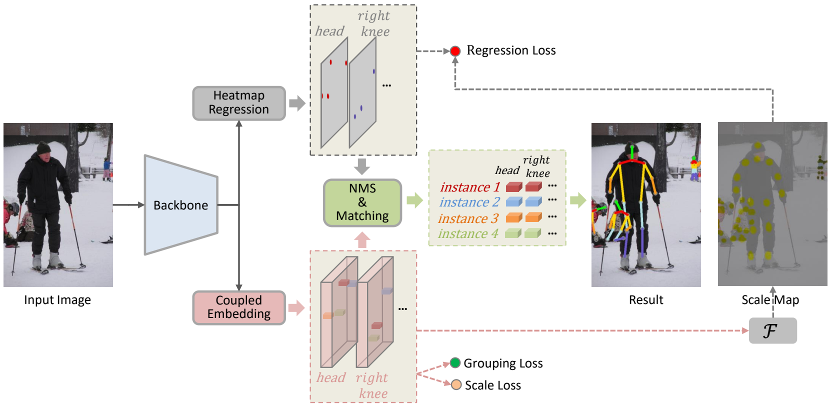


Fig. 3. An overview of our model. For each body keypoint, the model simultaneously predicts detection heatmaps (gray) and embedding maps (pink). Function $\mathcal{F}$ tranforms embedding maps into scale maps. Each location in scale maps represents a scale factor. We draw the valid scale factors with yellow color, where larger value corresponds to higher brightness. To make it easier to understand, locations in scale maps with low detection scores are masked.

Figure 3 illustrates the framework of our model. There are two branches that produce detection heatmaps and embedding maps, respectively. During training, regression loss is employed to supervise detection heatmaps and grouping loss is employed to supervise the identity embeddings. The scale loss acts as a regularization item to constrain the embeddings. In addition, the vectors in embedding maps can be transformed into scale factors which adaptively adjust the ground truth heatmaps of detection branch. For a fair comparison with the baseline method, we choose the same post processing as AE [22]. We first utilize non-maximum suppression to get the peak detections for each keypoint. Then we retrieve their corresponding embeddings at the same pixel location in embedding maps. At last, we identify the group of detected keypoints across body parts by matching embeddings via Hungarian algorithm [15]. To verify the generality of our method, we further apply our method to HigherHRNet [4].

3.2 Coupled Embedding

Coupled Embedding predicts embedding maps $\boldsymbol{T} \in \mathcal{R}^{H \times W \times K \times M}$ for K types of keypoint, where M denotes the dimensionality of embeddings, H and W denotes the height and width of embedding maps, respectively. Following AE [22], we adopt the grouping loss to prompt embeddings within an instance tend to have close distance, while embeddings across instances tend to be far apart. At the same time, we expect that the multi-dimensional embeddings can be sparse. Hence, we attempt to constrain the embeddings so that the principal components of embeddings are concentrated in certain dimensions, while keeping the values in other dimensions small. Based on the fact that the scales of instances usually have significant variations, we utilize the scales to generate vectors for embedding regularizing. In this paper, we define the scale of instance as $S_n = \sqrt{S_{box_n}}/L$, where S_{box_n} denotes the bounding box area of the n^{th} instance and L denotes the short size of the input image.

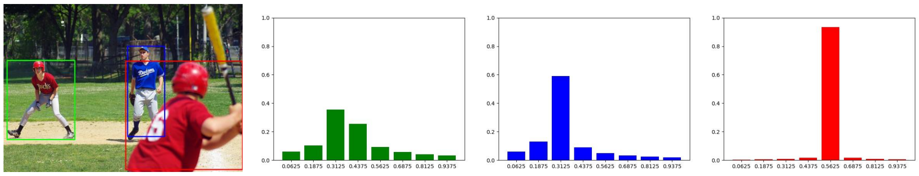


Fig. 4. Regularization vectors of three different instances in an image. The x-axis and y-axis denote scale level and the normalized component value, respectively.

Suppose $\boldsymbol{s}_n$ denotes the generated regularization vector of the n^{th} instance, we compute $\boldsymbol{s}_n$ as:

$$\boldsymbol{s}_n = \frac{(|S_n - \boldsymbol{d}|)^{-1}}{\|(|S_n - \boldsymbol{d}|)^{-1}\|_2}, \tag{1}$$

where S_n denotes the scale of the n^{th} instance and the $\boldsymbol{d}$ is a constant vector which divides the normalized scale into multiple levels. The value of each dimension of $\boldsymbol{d}$ and can be written as:

$$d_i = \frac{0.5 + i}{M}(i = 0, 1, \cdots, M - 1) \tag{2}$$

There is an example of regularization vectors of three different instances in Figure 4. As shown in Figure 4, the principal components of different regularization vectors are different. The dimension with the highest value corresponds the scale level which is the closest to the scale of instance.

3.3 Improving Heatmap Regression with Coupled Embedding

The 2D gaussian activation functions of heatmap regression in previous methods are totally similar, which can be written as:

$$\boldsymbol{h}_{n,k,i,j} = e^{\frac{-((i-x_{n,k})^2+(j-y_{n,k})^2)}{2\sigma^2}}, \tag{3}$$

where $\{x_{n,k}, y_{n,k}\}$ denotes the coordinate of each keypoint and $\{i, j\}$ denotes the coordinate of each pixel in heatmap. σ is the standard deviation of gaussian kernel which is usually set as a constant in previous methods. However, fixing standard deviation is unreasonable because the model needs to handle a large variance of human scales. An intuitive approach to solve this problem is applying a scale factor to adaptively adjust the standard deviation of gaussian kernel, which can be written as:

$$\boldsymbol{h}^{\gamma}_{n,k,i,j} = e^{\frac{-((i-x_{n,k})^2+(j-y_{n,k})^2)}{2(\sigma\cdot\gamma_n)^2}} = (\boldsymbol{h}_{n,k,i,j})^{1/\gamma_n^2}, \tag{4}$$

where γ_n denotes scale factor. And we define the scale factor as $\gamma_n = S_n/\theta$, where θ is a hyper-parameter to adjust the range of scale factor value.

Table 1. Comparison of training with fixed standard deviation and training with adjusted standard deviation. We apply associative embedding[22] as the baseline.

σ	fixed	adjusted					
θ	-	0.85	0.75	0.65	0.55	0.45	0.35
AP	64.8	59.8	60.8	63.9	61.6	61.4	61.7

We attempt to directly take the scale of instance as the scale factor to adjust the standard deviation, however we find it is inferior to prior fixed standard deviation. We conduct multiple experiments in different values of θ, and the results is shown in Table 1. We argue that the reason of the precision degradation is that complex pose, occlusion and partially labeling cause the scale of instance contains much noise. To tackle this problem, we utilize learned embedding to generate scale factor which eliminates the influence of scale noise and can been written as:

$$\gamma_{i,j} = \frac{1}{\theta} \sum_{m=0}^{M-1} \hat{t}_{i,j,m} d_m, \tag{5}$$

where $\hat{\boldsymbol{t}}_{i,j}$ is the normalized value of $|\boldsymbol{t}_{i,j}|$ and $\boldsymbol{d}_m$ is the constant vector defined in Equation 2.

3.4 Loss Function

For Coupled Embedding, we simultaneously impose regularization loss and grouping loss on embedding maps. We denote regularization loss and grouping loss as $\mathcal{L}_s$ and $\mathcal{L}_g$, respectively. The regularization loss is written as:

$$\mathcal{L}_s = \frac{1}{NK} \sum_n \sum_k (1 - \langle \hat{\boldsymbol{t}}_{n,k}, \boldsymbol{s}_{n,k} \rangle), \tag{6}$$

where $\hat{\boldsymbol{t}}_{n,k}$ is sampled in embedding maps and $\boldsymbol{s}_{n,k}$ is the corresponding regularization vectors. The $\mathcal{L}_s$ maximizes the cosine similarity between $\hat{\boldsymbol{t}}_{n,k}$ and $\boldsymbol{s}_{n,k}$. Grouping loss encourages pairs of embeddings to be assigned similar values if the corresponding keypoints belong to the same instance or dissimilar values otherwise. The loss function can be written as:

$$\mathcal{L}_g = \frac{1}{NK} \sum_n \sum_k \|\boldsymbol{t}_{n,k} - \bar{\boldsymbol{t}}_n\|_2^2 + \frac{2}{N(N-1)} \sum_n \sum_m e^{-\|\bar{\boldsymbol{t}}_n - \bar{\boldsymbol{t}}_m\|_2^2/2}, \tag{7}$$

where $\bar{\boldsymbol{t}}_n = \frac{1}{K} \sum_k \boldsymbol{t}_{n,k}$.

In detection branch, distances between predicted heatmaps and target heatmaps are frequently measured by L2 loss [2,22,24,14]. In target heatmaps, background samples make up the vast majority, which leads to imbalanced training data. To tackle this problem, we adaptively decays the loss value of easy samples, which is similar to focal loss [19] in classification. In [16,32,21], this idea is also applied to improve heatmap regression. The regression loss can be written as:

$$\mathcal{L}_r = \boldsymbol{W} \cdot \|\boldsymbol{H}_p - \boldsymbol{H}_g^{\sigma \cdot \Gamma}\|_2^2, \tag{8}$$

where $\boldsymbol{H}_p$ is the predicted heatmap and $\boldsymbol{H}_g^{\sigma \cdot \Gamma}$ is the adaptively adjusted ground truth. $\boldsymbol{W}$ is the weight which can be defined as:

$$\boldsymbol{W} = \boldsymbol{P} \cdot |1 - \boldsymbol{H}_p| + (1 - \boldsymbol{P}) \cdot |\boldsymbol{H}_p|, \tag{9}$$

$$\boldsymbol{P} = (1 - log^{\boldsymbol{H}_g^{\sigma \cdot \Gamma}})^{-\beta}, \tag{10}$$

where β is the hyper-parameter that controls the decay rate of the noncentral sample. In our practice, we set β to 0.01. The $\boldsymbol{P}$ defines the likelihood that a sample is positive. Equation (9) shows that positive samples which predict high activations are assigned low weights and negative samples which predict low activations are assigned low weights.

In conclusion, the total loss of our method can be written as:

$$\mathcal{L}_{total} = \mathcal{L}_r + \lambda_1 \mathcal{L}_g + \lambda_2 \mathcal{L}_s, \tag{11}$$

where λ_1 and λ_2 are hyper-parameters. In practice, we set λ_1 to 1e-3 and λ_2 to 1e-4.

4 Experiments

4.1 Datasets and Implementation Details

Datasets In this paper, we validate our method on CrowdPose [18] and COCO Keypoint [20] benchmarks. CrowdPose consists of three splits of *Train*, *Val*, *Test* with 10K, 2K, 8K images, respectively. COCO Keypoint has larger scale and includes three splits of *Train*, *Val*, *Test-dev* with 64K, 5K, 20K images, respectively. The images in COCO Keypoint are usually collected from daily life where crowded scenes only account for a small portion. CrowdPose defines an index to represent the crowding level of input images and sets up the benchmark that covers various scenes.

Implementation Details Our models are trained with Adam optimizer [13] for a total of 300 epochs on both COCO [20] and CrowdPose [18]. The base learning rate is initialized to 1.5e-3 at the beginning, then dropped to 1.5e-4 and 1.5e-5 at the 200^{th} and 260^{th} epochs, respectively. Following previous works [22,4,21], we apply data augmentation with random scale ($[0.75, 1.5]$), random rotation ($[-30°, 30°]$), random translation ($[-40, 40]$) and random horizontal flipping in a probability of 0.5. During test, the short side of input image is resized to 512 (or 640) and the flip test is also performed in all experiments, which is the same as previous works [22,4,21]. When we perform multi-scale test, we resize the original image with multiple scale factors which are set to {0.5, 1.0, 1.5}.

4.2 Comparison with SOTA

CrowdPose Test We compare our Coupled Embedding (CE) with state-of-the-art HPE methods on CrowdPose Test whose results are shown in Table 2. Our method achieves the best performance among all methods for both single and multi-scale test. Top-down methods do not perform well in crwoded scenes and get lower AP scores than bottom-up methods. Compared with the SOTA embedding-based method [21] (this method is based on HigherHRNet [4]), when the model is not pre-trained on COCO, our method achieves 1.8 points gain in AP and 1.6 points gain in AP^H (highly crowded scenes). After pre-training on COCO, the AP gain reduces to 0.5 but our method still achieves 1.2 points gain in AP^H. Compared with other bottom-up methods, the superiority of our method is more apparent. Overall, the strong results on CrowdPose Test especially for AP^H demonstrate that our method is excellent at handling images in crowded scenes.

Table 2. Comparisons on CrowdPose Test. Superscripts E, M and H of AP stand for easy, medium and hard. Superscript * means multi-scale test. Subscript $_\dagger$ means model is pretained on COCO

Method	AP	AP^{50}	AP^{75}	AP^E	AP^M	AP^H
Top-down methods						
Mask-RCNN[10]	57.2	83.5	60.3	69.4	57.9	45.8
AlphaPose[5]	61.0	81.3	66.0	71.2	61.4	51.1
SimpleBaseline[30]	60.8	84.2	71.5	71.4	61.2	51.2
SPPE[18]	60.0	84.2	71.5	75.5	66.3	57.4
Bottom-up methods						
OpenPose[2]	-	-	-	62.7	48.7	32.3
HigherHRNet-W48[4]	65.9	86.4	70.6	73.3	66.5	57.9
HigherHRNet-W48*[4]	67.6	87.4	72.6	75.8	68.1	58.9
DEKR-W32[8]	65.7	85.7	70.4	73.0	66.4	57.5
DEKR-W48[8]	67.3	86.4	72.2	74.6	68.1	58.7
DEKR-W48*[8]	68.0	85.5	73.4	76.6	68.8	58.4
SWAHR-W32[21]	66.7	86.9	71.7	74.3	67.3	58.9
SWAHR-W48[21]	68.0	88.1	72.9	75.2	68.5	60.5
SWAHR-W48*[21]	69.7	89.0	75.1	77.2	70.4	61.6
CE-W32(HRNet)	68.9	89.0	74.2	76.3	69.5	60.8
CE-W48(HRNet)	70.1	89.8	75.5	77.5	70.8	62.2
CE-W32(HigherHRNet)	69.6	89.7	74.9	76.9	70.3	61.6
CE-W48(HigherHRNet)	70.5	89.9	76.0	77.7	71.1	62.4
CE-W48*(HigherHRNet)	**71.6**	**90.1**	**77.3**	**79.0**	**72.2**	**63.3**
Bottom-up methods pre-trained on COCO						
SWAHR-W48$_\dagger$[21]	71.6	88.5	77.6	78.9	72.4	63.0
SWAHR-W48$^*_\dagger$[21]	73.8	90.5	79.9	81.2	74.7	64.7
CE-W48$_\dagger$(HigherHRNet)	72.9	89.5	78.8	79.6	73.7	64.5
CE-W48$^*_\dagger$(HigherHRNet)	**74.5**	**91.1**	**80.2**	**81.3**	**75.4**	**66.2**

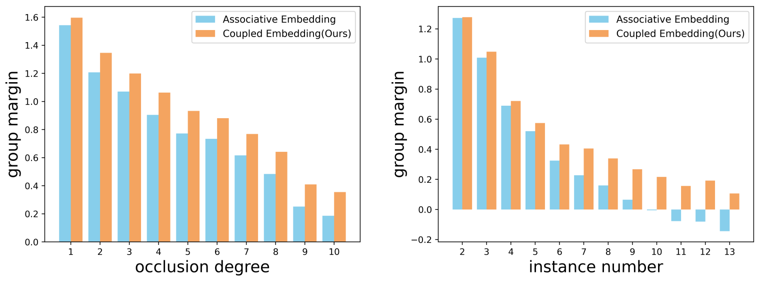
COCO Test-dev As shown in Table 3, we make comparisons with the state-of-the-art HPE methods on COCO Test-dev which is dominated by top-down methods. Compared with the performance on CrowdPose Test, our method has lower AP gain on COCO Test-dev since there are less images in complex scenes. However, our method still outperforms other bottom-up methods. And when we evaluate the model with multi-scale test, we can get the highest AP score at 72.8. This achieves a new state-of-the-art result. Compared with top-down methods, our best result has an advantage in AP score over early methods [10,25,3] and is comparable with recent top-down methods [3,30].

Table 3. Comparisons on COCO Test-dev. AE(HRNet-W32) means a implemention of associative embedding in [4], where the model replaces the backbone with HRNet. Superscripts M and L of AP stand for medium and large. Superscript * means multi-scale test

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Top-down methods					
Mask-RCNN[10]	63.1	87.3	68.7	57.8	71.4
G-RMI[25]	64.9	85.5	71.3	62.3	70.0
CPN[3]	72.1	91.4	80.0	68.7	77.2
SimpleBaseline[30]	73.7	91.9	81.1	70.3	80.0
HRNet-W48[28]	75.5	92.5	83.3	71.9	81.5
Bottom-up methods					
OpenPose[2]	61.8	84.9	67.5	57.1	68.2
AE(HRNet-W32)[4]	64.1	86.3	70.4	57.4	73.9
PersonLab[24]	66.5	88.0	72.6	62.4	72.3
PersonLab*[24]	68.7	89.0	75.4	64.1	75.5
PifPaf[14]	66.7	-	-	62.4	72.9
SPM[23]	66.9	88.5	72.9	62.6	73.1
HigherHRNet-W32[4]	66.4	87.5	72.8	61.2	74.2
HigherHRNet-W48[4]	68.4	88.2	75.1	64.4	74.2
HigherHRNet-W48*[4]	70.5	89.3	77.2	66.6	75.8
DEKR-W32[8]	67.3	87.9	74.1	61.5	76.1
DEKR-W48[8]	70.0	89.4	77.3	65.7	76.9
DEKR-W48*[8]	71.0	89.2	78.0	67.1	76.9
SWAHR-W32[21]	67.9	88.9	74.5	62.4	75.5
SWAHR-W48[21]	70.2	89.9	76.9	65.2	77.0
SWAHR-W48*[21]	72.0	90.7	78.8	67.8	77.7
CE-W32(HRNet)	67.0	88.9	73.7	60.4	76.4
CE-W48(HRNet)	68.4	88.7	75.5	63.8	75.9
CE-W32(HigherHRNet)	68.8	90.3	75.2	62.9	77.1
CE-W48(HigherHRNet)	71.1	90.8	77.8	66.4	78.0
CE-W48*(HigherHRNet)	**72.8**	**91.2**	**79.9**	**68.3**	**79.3**

4.3 Group Margin

In order to measure the keypoint grouping competence of embedding-based methods without AP, we introduce an index named group margin which is defined as the minimum embedding distance minus grouping threshold. Then we evaluate images on CrowdPose Test [18] and collect group margins of each image for above two methods. At last, we calculate the mean group margin of test images.



(a) (b)

Fig. 5. (a) shows the group margins in 10 different occlusion degrees. (b) shows the group margins correspoding to number of instances from 2 to 13. In (b), the negative group margin indicates the embedding distance is lower than grouping threshold.

Comparing the group margin in Figure 5(b), we observe that AE [22] is sensitive to number variation of instances in an image. When the number of instances is higher than ten, the embedding distance is lower than grouping threshold. Compared with AE [22], our method achieves larger group margin in each occlusion degree and number of instances. It indicates our model has more powerful grouping capacity in crowded scenes. In Figure 5(a), we observe that our advantages in different occlusion degrees are relatively stable. In Figure 5(b), our advantage is overwhelming when number of instances is large. Large number of instances in image means large number of clustering centers in embedding space. This experiment result illustrates that if the number of clustering centers is large, some embeddings could be assigned to the wrong group, where the embeddings of different instances are likely to be linearly inseparable.

4.4 Comparison of Keypoint Grouping

Our method has obvious advantages in keypoint grouping, but it only indicates in some complex scenes such as high occlussion degree or large number of instances. However, many test samples are in simple scenes in our test sets. Hence, to explore the superiority of our method in keypoint grouping, we sample 4 subsets from the CrowdPose Test according to different occlussion degrees and instance numbers. To eliminate the influence of keypoint detection, all methods in the experiments apply naive heatmap regression.

All the results are shown in Table 4. Subset1 has small number of instances but high occlusion degree. In this scene, embedding-based methods indiacate obvious advantage and our method has larger superiority. Subset2 has small number of instances and low occlusion degree, which is a commom simple scene.

Table 4. Comparisons on 4 different subsets. We apply 3 methods in this experiment. DEKR belongs to dense regression methods and AE [22] is the vanilla embedding-based method. The backbone and the input resolution are the same in all the methods.

		subset1	subset2	subset3	subset4
number of instances		$2 \sim 5$	$2 \sim 5$	> 6	> 6
occlusion degree		top 20%	last 20%	top 20%	last 20%
AP	DEKR	61.8	78.7	50.3	**62.5**
	AE	64.4	**78.8**	49.8	60.3
	ours(AE + scale loss)	**65.9**	78.7	**52.0**	62.2

In this scene, all the methods achieve similar AP scores. Subset3 has high occlusion degree and large number of instances which is a pretty complex scene. In this scene, our method performs much better than other methods. Subset4 has low occlusion degree but large number of instances. In this scene, vanilla embedding-based method has noticeable performance degradation but our method alleviates this negative influence. The above results show the superiority of our method in keypoint grouping when the occlusion degree is high or number of instances is large.

4.5 Ablation Study

Table 5. Ablation study of different strategies. We apply AE [22] with the backbone of HRNet-W32 as our baseline.

baseline	√	√	√	√	√
embedding regularization		√	√	√	√
adjusted heatmap				√	√
weighted L2-loss			√		√
AP	64.8	66.0	67.6	67.6	**68.9**
AP^{hard}	57.5	58.2	59.6	59.8	**60.8**

We perform a series of ablation experiments on CrowdPose Test to study the effects of our strategies. As we can see in Table 5, all of our designs lead to obvious increases in AP. Eventually, our method gets around +4 AP gain over the baseline. Compared to the baseline, the embedding regularization achieve a gain of +1.2 AP. The adjusting for ground truth heatmaps and weighted L2-loss brings both bring apparent improvement. As keypoint grouping is based on the results of keypoint detection, we utilize these two strategies to improve heatmap regression for a fair comparison with current SOTA embedding-based methods. Besides mean AP, all of our strategies bring improvements on AP^{hard}.

4.6 Hyper-parameter Study

Table 6. Study of M and θ. The results are reported on CrowdPose Test and COCO Val with single-scale test. Detection heatmap in each model is trained by standard L2-loss.

(a) $\theta = 0.55$

M	4	8	12	16
AP	67.1	**67.6**	67.6	67.3

(b) $M = 8$

	CrowdPose Test					COCO Val				
θ	0.45	0.50	0.55	0.60	0.65	0.30	0.35	0.40	0.45	0.50
AP	66.8	67.5	**67.6**	67.2	66.9	65.1	66.1	**66.4**	66.0	61.8

For Coupled Embedding, we study two significant hyper-parameters in Equation 5 including M and θ which denote the dimensionality of embeddings and the coefficient of scale factors, respectively. As shown in Table 6(a), we can get the best result when M is set to 8 or 12. Considering both efficiency and accuracy, we finally set M to 8. The θ is a coefficient that controls the scaling amplitudes of scale factors. To explore an appropriate θ, we perform experiments with different values of θ on CrowdPose and COCO. Considering results in Table 6(b), we set θ to 0.55 and 0.4 when we experiment on CrowdPose and COCO, respectively.

5 Conclusions

In this paper, we probe the limitation of keypoint grouping in Associative Embedding [22] and find that different dimensions of embeddings in an image are highly linearly correlated. To address this issue, we propose a novel keypoint grouping method named Coupled Embedding for bottom-up human pose estimation. Our method imposes an additional constraint on embeddings to learn sparse multidimensional embeddings. Our method creates new state-of-the-art results on CrowdPose Test (74.5 AP) and COCO Test-dev (72.8 AP), outperforming all existing bottom-up methods. We conduct a series of experiments and the results show that our model has significant advantages in complex scenes.

Acknowledgments: This work was supported by Key-Area Research and Development Program of Guangdong Province (No.2019B010153001). This work is being sponsored by Zhejiang Lab (No. 2021KH0AB07). This work was also supported by National Natural Science Foundation of China under Grants 62006230, 62076235.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3686–3693 (2014)
2. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Real-time multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018)
3. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7103–7112 (2018)
4. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5386–5395 (2020)
5. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2353–2362 (2017)
6. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems 26. vol. 26, pp. 2121–2129 (2013)
7. Frome, A., Singer, Y., Sha, F., Malik, J.: Learning globally-consistent local distance functions for shape-based image retrieval and classification. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8 (2007)
8. Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. arXiv preprint arXiv:2104.02300 (2021)
9. Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., Lazebnik, S.: Improving image-sentence embeddings using large weakly annotated photo collections. In: 13th European Conference on Computer Vision, ECCV 2014. pp. 529–545 (2014)
10. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 386–397 (2020)
11. Huang, P., Huang, Y., Wang, W., Wang, L.: Deep embedding network for clustering. In: ICPR '14 Proceedings of the 2014 22nd International Conference on Pattern Recognition. pp. 1532–1537 (2014)
12. Iqbal, U., Gall, J.: Multi-person pose estimation with local joint-to-person associations. In: European Conference on Computer Vision. pp. 627–642 (2016)
13. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. In: ICLR 2015 : International Conference on Learning Representations 2015 (2015)
14. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11969–11978 (2019)
15. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly **2**(1), 83–97 (1955)
16. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 734–750 (2018)
17. Li, J., Su, W., Wang, Z.: Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11354–11361 (2020)

18. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10863–10872 (2019)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(2), 318–327 (2020)
20. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755 (2014)
21. Luo, Z., Wang, Z., Huang, Y., Tan, T., Zhou, E.: Rethinking the heatmap regression for bottom-up human pose estimation. arXiv preprint arXiv:2012.15175 (2020)
22. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: 31st Annual Conference on Neural Information Processing Systems, NIPS 2017. vol. 30, pp. 2278–2288 (2017)
23. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6951–6960 (2019)
24. Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. Springer, Cham (2018)
25. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3711–3719 (2017)
26. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4929–4937 (2016)
27. Sun, K., Geng, Z., Meng, D., Xiao, B., Liu, D., Zhang, Z., Wang, J.: Bottom-up human pose estimation by ranking heatmap-guided adaptive keypoint estimates. arXiv preprint arXiv:2006.15480 (2020)
28. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5693–5703 (2019)
29. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research **10**(9), 207–244 (2009)
30. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 472–487 (2018)
31. Yang, B., Fu, X., Sidiropoulos, N.D., Hong, M.: Towards k-means-friendly spaces: simultaneous deep learning and clustering. In: ICML'17 Proceedings of the 34th International Conference on Machine Learning - Volume 70. pp. 3861–3870 (2017)
32. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)