Occlusion-Aware Siamese Network for Human Pose Estimation

Lu Zhou[1,2], Yingying Chen[1,2,3], Yunze Gao[1,2], Jinqiao Wang[1,2,4], and Hanqing Lu[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China
[3] ObjectEye Inc., Beijing, China
[4] NEXWISE Co., Ltd., Guangzhou, China
{lu.zhou,yingying.chen,yunze.gao,jqwang,luhq}@nlpr.ia.ac.cn

Abstract. Pose estimation usually suffers from varying degrees of performance degeneration owing to occlusion. To conquer this dilemma, we propose an occlusion-aware siamese network to improve the performance. Specifically, we introduce scheme of feature erasing and reconstruction. Firstly, we utilize attention mechanism to predict the occlusion-aware attention map which is explicitly supervised and clean the feature map which is contaminated by different types of occlusions. Nevertheless, the cleaning procedure not only removes the useless information but also erases some valuable details. To overcome the defects caused by the erasing operation, we perform feature reconstruction to recover the information destroyed by occlusion and details lost in cleaning procedure. To make reconstructed features more precise and informative, we adopt siamese network equipped with OT divergence to guide the features of occluded images towards those of the un-occluded images. Algorithm is validated on MPII, LSP and COCO benchmarks and we achieve promising results.

Keywords: siamese network, occlusion, human pose estimation

1 Introduction

2D human pose estimation has enjoyed great success in recent years owing to the development of deep neural networks. It aims at predicting the positions of human joints given a single RGB image and serves as a significant basis for several vision tasks such as action recognition [38, 26], person re-identification [27] and human-computer interaction [20]. Nevertheless, it is still confronted with a lot of challenges such as view changes, complex human gestures, human joints scale changes and occlusion. Among these troublesome factors, occlusion shown in Fig. 1 poses great degradation to the performance of human pose estimation. In general, presentation of occlusions leads to contamination of deep features and confuses the network to make incorrect decisions.

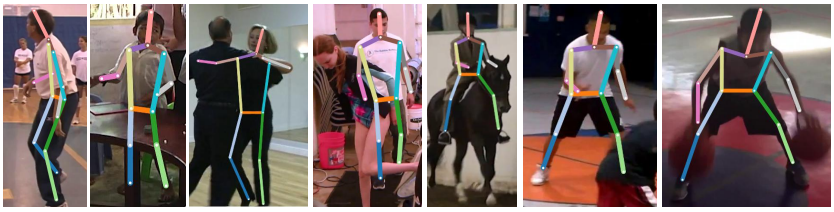

Fig. 1. Illustration of the results on some occluded images of MPII datasets.

Previous methods based on deep neural networks [28, 21, 36, 3] mainly focused on searching more efficient and powerful architectures of the deep neural networks. [8] utilized CRF-based attention to settle the occlusion problem. [4] proposed an adversarial posenet to tackle this issue. However, attention used in [8] leads to large increases of the parameters and computation cost while adversarial network in [4] is usually hard to converge. In this paper, an occlusion-aware siamese network is proposed to conquer the dilemma caused by occlusion and surpasses these mentioned methods based on the same backbone.

We leverage the attention mechanism to exclude the interference of occlusions. Among those popular human pose estimation benchmarks, COCO keypoint detection dataset [16], MPII dataset [1], usually, occlusion flag is offered as another labeling information. Here, we employ the occlusion flag which is seldom excavated before to predict corresponding occlusion circumstances and obtain occlusion-aware attention map. Compared with previous attention mechanisms [8, 42] used in human pose estimation, attention map employed here is learned explicitly via intermediate supervision, which is more purposeful and can predict occlusion more precisely. The obtained attention map serves as a solid foundation for feature erasing and reconstruction.

Feature erasing means erasing the contaminated feature map and provides cleaner representation. Guided by explicitly learned occlusion-aware attention map, we can remove the ambiguities caused by occlusion and obtain relatively cleaner feature map.

However, cleaned feature map cannot provide precise and holistic description of the whole human skeleton due to the missing semantic information. The erasing procedure not only deletes tremendous incorrect expressions but also mistakenly gets rid of some informative cues especially under the circumstances of self-occlusion. Hence, feature reconstruction is necessary for obtaining more powerful and informative feature representations. On one hand, feature reconstruction attains refreshed information to replace those occluded features in the case that semantics are destroyed by occlusion. On the other hand, feature reconstruction is able to recover the wrongly removed semantics when useful information is mistakenly erased. Here, we design a feature reconstruction submodule which can capture information from surrounding areas without occlusion to facilitate the recovering.

To provide ample prior guidance for reconstruction, here, we advance a siamese framework to facilitate this process. The siamese network possesses two branches overall where model weights are shared between them. The second branch takes images without artificial occlusion as input to extract the clean feature representation. In contrast, the first branch takes occluded images which contain the same content as the second branch except for the appearance of the occlusion as input. The occlusion appearing in the first branch is manually created. Purpose of the siamese structure is to make the occluded branch imitate the behavior of the branch without occlusion. How to pull these two branches more closely in high-dimension feature space is challenging. In this paper, we employ optimal transport (OT) divergence [18, 25] with additional mask as deluxe regularization instead of correspondence-based approaches to achieve this purpose. Branch without occlusion encodes more confident information whilst the other branch is less confident somehow. The introduction of OT divergence enables the less confident feature to be aligned with the more confident one.

Integrating submodules aforementioned together forms our occlusion-aware siamese network and whole framework is named as OASNet which aims at settling the occlusion problem. The main contributions are three folds:

- We propose a feature erasing and reconstruction submodule to obtain cleaner feature representation and reconstruct the erased feature. Different from previous methods, attention map which is occlusion-aware is put forward to remove the ambiguities caused by different types of occlusion.
- To make the occluded feature mimic the behavior of feature without occlusion, siamese network equipped with erasing and reconstruction submodule is put forward.
- Instead of utilizing correspondence-based methods to narrow the discrepancy of these two sets features, we adopt the optimal transport to fulfill this task. Our model makes the first attempt to perform human pose estimation under the primal form of optimal transport algorithm and proves effective.

2 Related Work

Human Pose Estimation. Recent years human pose estimation has achieved promising progress due to the application of deep neural networks. DeepPose [34] made the first attempt to integrate human pose estimation and deep neural network together. Pose estimation was regarded as a regression task and coordinates of human joints were predicted directly as model output. Nevertheless, directly regressing human joints was somewhat difficult and heatmap regression methods [6, 21, 36, 8, 39, 30, 13, 3, 37] emerged as a new fashion. Among these methods, [33, 32, 11, 19] were devoted to reducing the false positive predictions via different types of Markov Random Field (MRF). [7] concentrated on building the spatial relationships of different human joints features with the Conditional Random Field (CRF). In addition to these deep graphical models, searching more efficient network structures yielded state-of-the-art performance. [36] proposed the concept of deep convolutional pose machines to enlarge the respective field

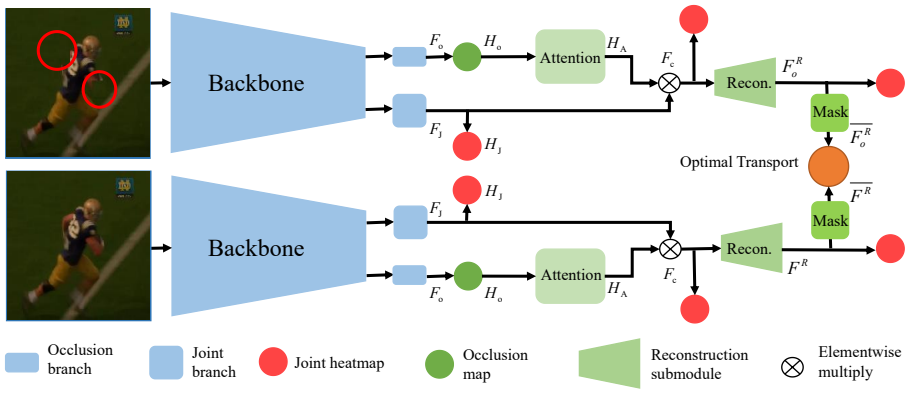

Fig. 2. Illustration of proposed occlusion-aware siamese network. There exist two branches where the first branch takes the artificially occluded image as input and the second branch takes the primal image as input.

and refined initial predictions via cascaded design. Hourglass [21], also called conv-deconv structure, adopted the cascaded design as well to refine previously generated results step by step. Subsequent works such as [8, 39, 30, 13, 22, 40, 29] all took hourglass as their backbone to extract efficient feature representations and implemented their ideas. [28] proposed HRNet which aimed at maintaining high resolution features across all stages of the network and achieved state-of-the-art performance across several human pose benchmarks.

3 Method

Overview of the proposed framework can be found in Fig. 2. In this section, we will detail the proposed occlusion-aware siamese network in three aspects, feature erasing and reconstruction, siamese framework and optimal transport divergence.

3.1 Feature Erasing and Feature Reconstruction

Occlusion poses great threat to human pose estimation due to missing semantics of corresponding human joints and contamination of the features. Existence of the occlusion confuses the network to make right decisions and hesitate around the area of occlusion. Hence, predicting occlusion flag serves as a necessary task to perceive the occlusion and provides essential cues for subsequent recognition and location. Previous methods generally take use of auto-learned attention to highlight the informative patches yet occluded regions are also picked out. It is difficult to perform the cleaning schedule in this case. In this paper, we exploit the occlusion flag which is additionally labeled to learn the attention explicitly. Previous works seldom leverage this labeling resource and ignore this valuable cue.

Fig. 3. Ground truth of occlusion heatmaps and corresponding predictions. Red rectangles demonstrate the predicted continuous occlusion patches.

To model this occlusion pattern, multi-task learning is adopted here. As shown in Fig. 2, predictions of occlusion map and joint heatmaps are regarded as two separate tasks and the whole prediction process is modeled as a multi-task learning framework. Hard sharing mechanism where backbone for feature extraction is shared is adopted here.

We try two different approaches to perform the multi-task learning. For the first one, we regard the prediction of occlusion status as a classification task,

$$L_o = -\sum_{i=1}^{N}\sum_{c=0}^{C'-1} p^{i,c} log(p^{i,c}), \tag{1}$$

where cross entropy loss is utilized, N represents the number of human joints and C' indicates class numbers. Value of C' is 3 where class 0 indicates "not labeled", 1 indicates "labeled but not visible" and 2 indicates "labeled and visible".

In addition, we also try modeling the occlusion prediction as a heatmap regression task. Supervision for heatmap regression is formulated as shown in Fig. 3. We sum the occluded joint heatmaps up and the resulted heatmap is clamped to $[0, 1]$. Loss for this branch is denoted as

$$L_o = ||H_o - \hat{H}_o||_2^2, \tag{2}$$

where H_o is the predicted occlusion map and $\hat{H}_o$ represents corresponding ground truth. From Fig. 3 we can observe that predicted occlusion heatmaps are capable of predicting the continuous occlusion patch especially under extremely severe occluded situations. Prediction manner shown in Equation 1 brings in performance drop and we abandon the utilization of it. Directly classifying occlusion flag seems to be a hard task. Loss for human joints regression is expressed as

$$L_J = \sum_{i=1}^{N} ||H_J^i - \hat{H}_J^i||_2^2, \tag{3}$$

where N represents the number of human joints. Overall loss for this multi-task learning framework is denoted as

$$L = L_J + \lambda_o L_o, \tag{4}$$

where λ_o is a hyper-parameter which modulates the proportion of losses with respect to these two tasks.

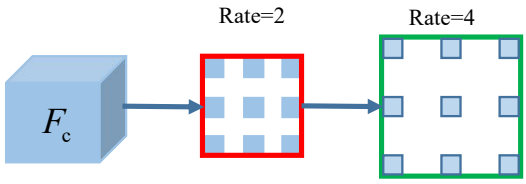

Fig. 4. Design of the reconstruction submodule. Two consecutive dilated convolution operations are adopted.

We convert the resulted occlusion heatmap to corresponding attention map via

$$H_A = 1 - H_o, \tag{5}$$

where $H_A \in R^{1\times H\times W}$ represents the occlusion-aware attention map. Operation of feature cleaning can be denoted as

$$F_c = F_J \odot H_A, \tag{6}$$

where feature map $F_J \in R^{C\times H\times W}$ is weighted by H_A. However, the missing information may deteriorate the overall performance. It is critical for the network to learn an automatic resuming mechanism and reconstruct the missing knowledge.

To obtain enough semantic information from surrounding areas of the occluded patches, two consecutive dilated convolutions with kernel 3×3 are adopted to ensure ample spatial coverage and enlarge respective field of the submodule. Dilation rates are 2 and 4 respectively. Maybe more complex design can exhibit better performance, however, it is not the main concern of this paper and the submodule here has shown promising advantage. Concrete demonstrations can be found in Fig. 4.

3.2 Siamese Framework

Feature reconstruction submodule reconstructs the erased features without any instruction. Though it takes advantage of context from neighboring regions, it still cannot recover the missing information precisely owing to the lack of prior guidance. The information recovered may not be adequate for final locating and still carries enormous distractive information. It is critical to provide a precise mentor for the reconstruction submodule to make the reestablishment process more reasonable. Perfect reestablishment may be the reproduction of the features of the images without occlusion. The reconstructed feature in this case can encode the same context as the un-occluded image.

It is intuitive to think of the teacher-student mode to perform the guidance procedure, i.e., distillation. Pre-trained model which takes the un-occluded images as input serves as the teacher and the model to be trained with occluded images acts as the student. Teacher model carrying more reliable sets of information provides more confident supervision for the student model. Nevertheless,

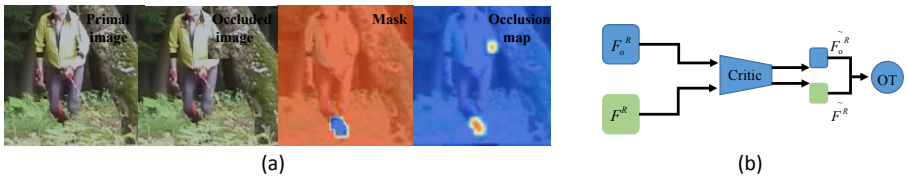

Fig. 5. (a) Illustrations of mask mechanism. (b) Illustrations of optimal transport submodule.

the features and parameters of pre-trained teacher model differ largely from the student and they are totally two different models. Directly executing the distillation over these two feature sets from two completely different models is challenging and not that adequate for our task.

In this paper, we abandon the usage of distillation and propose a novel siamese framework to conduct the mimicking process. As depicted in Fig. 2, siamese training is utilized to fulfill the mission of the similarity learning. The first branch takes the image with artificial occlusions as input and encodes the less confident information. The second branch takes the primal images without artificial occlusions and offers more confident information. The self-supervised training mechanism executes another form of teacher-student guidance and outperforms the distillation technique mentioned above.

To enforce artificial occlusions on primal images, we follow the same pasting operation as [13]. Random numbers of occlusion patches are chosen on each image with different scales to enrich the occlusion types. However, some primal images may have been endowed with some natural occlusions and it is non-trivial to perform the mimicking with mask to exclude the interference of the natural occlusions. The masked mimicking learning can effectively narrow the gap across the two branches of siamese network.

3.3 Optimal Transport Divergence

The most straightforward manner for reducing the gap between the two branches of siamese network is to narrow the corresponding pixel discrepancy. Various evaluation losses can be adapted into this kind of framework. Distinct from these methods, matching the distributions of these two branches acts as another substitute. In contrast, matching corresponding distributions demonstrates more superior advantages than correspondence-based techniques especially over the space of high dimension. In this section, we will detail the feature mimicking losses we explore.

Correspondence-based Approaches Correspondence-based methods seek to measure the discrepancy between these two branches by means of pixel correspondence. We denote the feature after reconstruction of the first branch as $F_o^R \in R^{C\times H\times W}$, and the feature of the second branch is expressed as $F^R \in R^{C\times H\times W}$.

Since there might exist some natural occlusions, it is necessary to exclude the guidance from these areas. In these areas, features from the second branch suffer from contamination as well and reconstruction may not be better than the first branch. We add a mask which excludes the mimicking of the areas with natural occlusions and is denoted as $\overline{M} \in R^{1\times H\times W}$. Illustrations can be found in Fig. 5(a). Feature map decorated by mask can be obtained via

$$\begin{aligned} \overline{F^R} &= F^R \odot \overline{M}, \\ \overline{F_o^R} &= F_o^R \odot \overline{M}, \end{aligned} \tag{7}$$

where $\overline{F^R}, \overline{F_o^R} \in R^{C\times H\times W}$ indicate masked feature maps. To minimize corresponding distinctions, $L2$ loss is adopted as follows:

$$L_{mimic} = \sum_{i=1}^{N} ||\overline{F^R}(i) - \overline{F_o^R}(i)||_2^2, \tag{8}$$

where i means the ith location and N is $H \times W$. Cosine similarity can be implemented following the similar way.

Distribution-based Approaches Aligning the distributions of these two groups features serves as another variant to eliminate the existing discrepancy. Usually KL divergence and JS divergence are two common measurements to assess the divergence between two distinct distributions. However, the comparison of these two distributions is still carried out on high-dimension space. To settle this dilemma, optimal transport divergence is adopted to compare the two distributions in a low-dimension feature space and relieves the difficulty.

Optimal transport seeks to find the optimal transportation strategy γ_0 to perform the mass moving from distribution u to distribution v and minimizes the transport cost. Usually, discretized version of optimal transport is adopted,

$$\begin{aligned} \gamma_0 &= \underset{\gamma\in\Pi(u,v)}{\arg\min} \langle \gamma, M\rangle_F, \\ \Pi(u,v) &\overset{def.}{=} \{\gamma \in R_+^{m_1\times m_2} : \gamma 1_{m_2} = u, \gamma^T 1_{m_1} = v\}, \end{aligned} \tag{9}$$

where $\langle ., . \rangle_F$ indicates the Frobenius dot product, $1_m := (1/m, ..., 1/m) \in R_+^m$. $M \in R_+^{m_1\times m_2}$ indicates the cost matrix.

Given the two sets of features F^R, F_o^R shown in Fig. 2, masked features are denoted as $\overline{F^R}, \overline{F_o^R}$ as well. Optimal transport divergence problem can be formulated as

$$\begin{aligned} L_{mimic} &= OT(\overline{F^R}, \overline{F_o^R}) \\ &= W_M(\overline{F^R}, \overline{F_o^R}) \\ &\overset{def.}{=} \min_{\gamma\in\Pi(u,v)} \langle \gamma, M\rangle_F, \end{aligned} \tag{10}$$

where $OT(\overline{F^R}, \overline{F_o^R})$ evaluates the distance between these two distributions. The channel number of $\overline{F^R}$ and $\overline{F_o^R}$ is denoted as C. At first, a critic which attempts

Algorithm 1 Iterative implementation of Sinkhorn divergence.

Input: Input masked feature maps $\overline{F^R}$ and $\overline{F_o^R}$, critic module D
Output: Sinkhorn loss $W_M(\overline{F^R}, \overline{F_o^R})$
1: Feature $\overline{F^R}$ and $\overline{F_o^R}$ are both sent into the critic, $\widetilde{F^R} = D(\overline{F^R})$ and $\widetilde{F_o^R} = D(\overline{F_o^R})$
2: $\forall(i,j)$ in $\widetilde{F^R}, \widetilde{F_o^R}$, $M_{ij} \overset{def.}{=} cos\_dis(\widetilde{F^R}(i), \widetilde{F_o^R}(j))$ and $M \in R^{C\times C}$
3: Initialize $b^{(0)} \leftarrow 1_C$
4: Compute Gibbs Kernel $K_{i,j} = exp(-M_{i,j}/\varepsilon)$
5:
for $r = 0; r < R; r++$ **do**
$a^{r+1} := \frac{1_C}{K^T b^r}, b^{r+1} := \frac{1_C}{K a^r}$
end for
6: Matrix P^R can be obtained via $diag(b^R) \cdot K \cdot diag(a^R)$
7: W_M can be obtained via: $\langle M, P \rangle$
8: return W_M

to down-sample these two groups features from $C \times H \times W$ to $C \times H^{'} \times W^{'}$ with only one convolution layer is utilized and channel dimension is kept unchanged. The changed features are denoted as $\widetilde{F^R}, \widetilde{F_o^R}$. Concrete illustrations can be found in Fig. 5(b). $\widetilde{F^R}, \widetilde{F_o^R}$ are then reshaped into $C \times k$ and $k = H^{'} \times W^{'}$. When it comes to the computation of W_M, we take the same policy as [15, 9] in an iterative manner. Sinkhorn divergence is hence implemented and reduces the computational complextiy. Cost M_{ij} is defined as the cosine distance. Illustrations of the algorithm can be found in Algorithm 1. Coefficient ε in Algorithm 1 is set to 0.1 and iteration number R is set to 5 finally.

Compared with correspondence-based manners, OT divergence overcomes the defect of sensitivity to the disturbances from outliers. Compared with KL divergence, for one hand, OT divergence conducts the similarity comparison in a low-dimension feature space. On the other hand, OT divergence relieves the issue of "vanishing gradient" under the circumstances of non-overlapped [2, 35] distributions and enforces more strict constraints. The optimal transport assures that distributions between F^R and F_o^R become closer and facilitates the mimicking process.

3.4 Training and Inference

Effectiveness of the algorithm is validated on two backbones Hourglass [21] and HRNet [28]. Erasing and reconstruction submodule is appended at the end of both two backbones.

Loss of the framework can be separated into three parts and denoted as follows:

$$L = L_J + \lambda_o L_o + \lambda_{mimic} L_{mimic}, \tag{11}$$

where $\lambda_o, \lambda_{mimic}$ represent the balancing factors to regulate the proportion of these losses and L_{mimic} represents all the feature mimicking losses mentioned in section 3.3. L_J and L_o take advantages of MSE loss which is a common scheme

in human pose estimation. Here, λ_o is taken as 1. Value of λ_{mimic} depends on the loss type. When L2 distance, cosine distance, and OT divergence are accepted, it is endowed with value 0.001. When KL divergence is taken, it is endowed with 100 to achieve better balance.

During inference, heatmaps fetched from the last stack of hourglass and the end of the HRNet are evaluated.

4 Experiments

4.1 Datasets

We evaluate our approach on three widely applied benchmarks, MPII [1], LSP [12] and COCO datasets [16]. MPII dataset contains 25k images with around 40k poses. LSP dataset consists of primal LSP split which contains 2k samples and LSP-extended split which possesses 10k samples. Usually, 1k samples out of the primal LSP dataset are used for test purpose and the other 11k samples are employed for training. COCO serves as one of the largest human pose benchmarks where 200k images are involved with 250k person instances labeled. In our implementation, 150k images with around 150k person instances are included during the training process. The proposed method is evaluated on the validation split where 5000 images are engaged.

4.2 Implementations

Data Augmentation. The experiments conducted on the MPII and LSP dataset take utilization of the same data augmentations scheme as previous works [39, 8]. During training, random rotation (± 30), color jittering, random scaling and random flipping are adopted. Input resolution is 256×256. Experiments on COCO dataset take advantage of completely the same data augmentations as [28] for fair comparison.

Training Schedule. All the experiments are conducted on the platform of Pytorch. For hourglass-based model, we employ RMSProp [31] to optimize the network. Learning rate is initialized with 5.0×10^{-4} and dropped by 10 at the epoch of 150, 170, and 200 with overall 220 epochs. For HRNet-based model, training schedule follows [28]. Learning rate for HRNet-based model is initialized with 1.0×10^{-3} and decayed by 10 at the epoch of 170 and 200. Adam optimizer [14] is adopted in HRNet-based model.

Test Schedule. Test over the MPII, LSP dataset follows [39, 8, 30] and takes use of six-scale pyramid pattern with flipping. Test over the COCO dataset follows the same procedure as [28] and bounding boxes used are kept the same as [28] as well.

4.3 Benchmark Results

MPII Dataset. Results on the MPII test set can be found in Table 1. Hourglass-based model (8 stacks) is used for test. We can observe that our approach

Table 1. Evaluation results using PCKh@0.5 as measurement on the MPII test set

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Insafutdinov et al. [10]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei et al. [36]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Newell et al. [21]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Ning et al. [23]	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2
Chu et al. [8]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Liu et al. [17]	98.4	96.4	92.0	87.9	90.7	88.3	85.3	91.6
Chou et al. [5]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Yang et al. [39]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke et al. [13]	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Tang et al. [30]	98.4	96.9	92.6	88.7	91.8	89.4	86.2	92.3
Sun et al. [28]	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
Zhou et al. [41]	98.5	96.9	92.8	89.3	91.8	89.5	86.4	92.5
Tang et al. [29]	98.7	97.1	93.1	89.4	91.9	90.1	86.7	92.7
ours (hg)	98.5	97.0	93.0	89.4	91.7	**90.3**	86.5	92.7
ours (HRNet-W32)	**98.8**	97.0	92.9	89.1	91.3	89.3	85.8	92.4

Table 2. Performance of our model based on HRNet-w32 over MPII validation split.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
HRNet	97.03	96.02	90.73	86.62	89.41	86.66	82.45	90.33
HRNet+Ours	97.24	96.50	91.07	86.81	89.55	87.08	83.63	90.71

achieves promising results compared with previous state-of-the-art methods. Final PCKh score is 92.7 and the improvement over baseline reaches up to 1.8%.

To validate the effectiveness of the algorithm, we try different backbones and results of HRNet-based methods over MPII validation split can be found in Table 2. For HRNet-based methods, we test on the MPII validation split with flipping and omit the six-scale pyramid testing for fair comparison. Compared with original HRNet which achieves 90.33 PCKh@0.5 score (We re-implement HRNet and the results is almost consistent with the results from official website.), our new algorithm achieves 90.71 PCKh@0.5 score and surpasses primal version by 0.38%. The occlusion-aware siamese network can still improve the performance even based on a strong baseline. Results over test split with 6-scale pyramid test can be found in Table 1, which achieve 92.4 PCKh score.

LSP dataset. Results of the LSP dataset can be found in Table 3. When training LSP dataset, MPII dataset is included following previous works [39, 8]. However, occlusion flag for LSP-extended split is omitted somehow and cannot provide precise supervision for LSP training. To obtain corresponding supervision for occlusion attention map, attention map from model pre-trained on MPII dataset is employed as the ground truth which serves as an approximation. From Table 3, we can observe that our model achieves promising results compared with previous state-of-the-art methods.

COCO dataset. For COCO dataset, evaluation results on the validation set are illustrated in Table 4. The model is based on primal HRNet and input size is set to 256×192. The results displayed of the other works are fetched from [28]. Following completely the same training and testing strategy as HRNet, our

Table 3. Evaluation results using PCK@0.2 as measurement on the LSP dataset

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Rafi et al. [24]	95.8	86.2	79.3	75.0	86.6	83.8	79.8	83.8
Insafutdinov et al. [10]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Wei et al. [36]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Chu et al. [8]	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Liu et al. [17]	98.1	94.0	91.0	89.0	93.4	95.2	94.4	93.6
Yang et al. [39]	98.3	94.5	92.2	88.9	94.4	95.0	93.7	93.9
Chou et al. [5]	98.2	94.9	92.2	89.5	94.2	95.0	94.1	94.0
Tang et al. [29]	98.6	95.4	93.3	89.8	94.3	95.7	94.4	94.5
ours (hg)	**98.8**	95.2	92.3	89.8	**95.2**	95.5	**94.7**	94.5

Table 4. Results on the COCO validation set.

Method	Backbone	Pretrain	Input Size	Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
HRNet-W32 [28]	HRNet-W32	N	256×192	28.5M	7.10	73.4	89.5	80.7	70.2	80.1	78.9
HRNet-W32 [28]	HRNet-W32	Y	256×192	28.5M	7.10	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W48 [28]	HRNet-W48	Y	256×192	63.6M	14.6	75.1	90.6	82.2	71.5	81.8	80.4
Ours	HRNet-W32	Y	256×192	29.9M	9.0	75.0	90.4	81.8	71.5	81.9	80.4
Ours	HRNet-W48	Y	256×192	66.0M	17.3	75.5	90.7	82.4	72.0	82.4	80.7

HRNet-w32 surpasses the baseline by 0.6 mAP and HRNet-w48 surpasses the baseline by 0.4 mAP.

Improvement over occluded and un-occluded human joints. Table 5 demonstrates the overall improvement over occluded and un-occluded human joints. The models are based on 2-stack hourglass model and HRNet. Improvement over occluded joints based on hourglass model is 1.72. The algorithm improves the performance over occluded joints. In contrast, the improvements over un-occluded joints is 1.12. These improvements can be ascribed to the reconstruction model which also benefits the un-occluded joints due to the capturing of more semantics. When utilizing HRNet as backbone, we can observe that improvement over occluded human joints makes the main contribution to the performance boosting. Progress over occluded human joints reaches about 1.51 PCKh score while progress over un-occluded human joints is only 0.19. From Table 5, we can observe that the algorithm improves the performance of occluded human joints across different backbones.

4.4 Ablation Study

We conduct the ablation study on the validation split of the MPII dataset and take 2-stack hourglass model whose result serves as the baseline as backbone. All the evaluation results are tested with single scale image and flipping operation is not involved.

Effectiveness of the E&R submodule. Different forms of the multi-task learning of occlusion prediction cause distinct effects. For classification-based methods shown in Eq. 1, performance drops obviously compared with baseline. In contrast, regression-based approach shown in Fig. 3 maintains the performance and

Table 5. Evaluation results using PCKh@0.5 as measurement on the MPII validation set with respect to the occluded joints and un-occluded joints

Method	Occluded	Un-occluded	Method	Occluded	Un-occluded
hourglass	72.59	92.06	HRNet	76.21	94.00
hourglass(ours)	74.31	93.18	HRNet(ours)	77.72	94.19

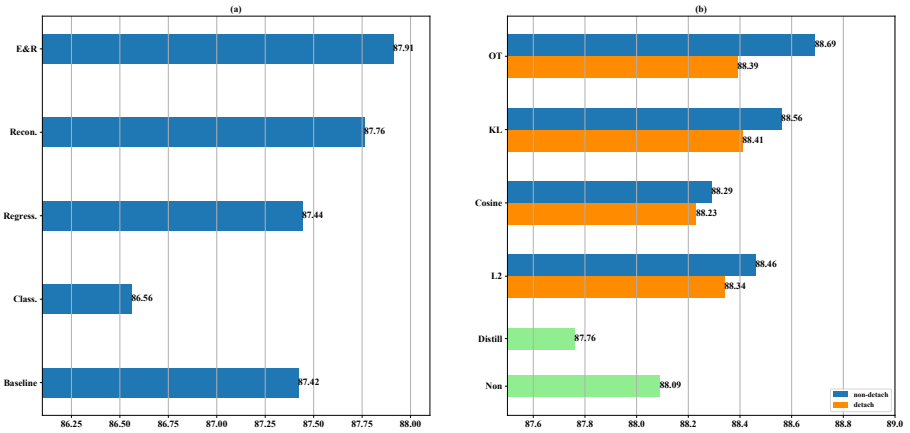

Fig. 6. (a) Effectiveness of the erasing and reconstruction submodule. "Class." means predicting occlusion flag via classification-based methods, "Regress." means predicting occlusion flag leveraging regression-based methods, "Recon." means the reconstruction without cleaning. "*E&R*" means cooperation of erasing and reconstruction. (b) Effectiveness of different losses of the feature mimicking. "L2","Cosine","KL","OT" are four different losses we adopted in our paper. "Non" represents the omitting of the mimicking loss. "detach" indicates operation of detachment of the un-occluded branch, whilst "non-detach" indicates the omitting of the detachment.

causes non-deterioration. The disparity may primarily come from the large gap of different types losses and we adopt the regression-based methods in the end. The effectiveness of the erasing and reconstruction submodule can be found in Fig. 6(a). We can note that reconstruction without cleaning improves the performance by 0.34%. The improvement mainly originates from the enrichment of context and enlargement of respective field, which confirms the rationality of the reconstructing design. When erasing operation is involved, the improvement promotes further. The erasing procedure excludes enormous interferences and thus benefits reconstruction process.

Effectiveness of the OT divergence. Effectiveness of different losses of the feature mimicking can be found in Fig. 6(b). For each loss listed in Fig. 6(b), we investigate two different formats of the mimicking. The first format means the feature of the un-occluded branch is detached to provide supervision for the occluded branch, yet the second format which is denoted as "non-detach" means that feature of the un-occluded branch is not detached and back propagated together with occluded branch to update the parameters of the network.

If L2 loss is adopted, PCKh score of the whole framework reaches up to 88.46 without "detach" operation and 88.34 with "detach" operation. If we utilize cosine distance, PCKh score reaches up to 88.23 with "detach" operation and 88.46 without "detach" operation. Both of these two methods improve the performance, which verifies the effectiveness of mimicking mechanism.

When KL divergence is applied as loss to narrow the gap between distributions, performance can be improved to 88.41 PCKh score with "detach" operation and 88.56 without "detach" operation. If optimal transport divergence is involved, performance can be further boosted up to 88.39 PCKh score with "detach" operation and 88.69 PCKh score without "detach" operation.

We can notice that distribution-based methods outperform correspondence-based approaches overall, which confirms our conjecture that narrowing the correspondence discrepancy over high-dimension feature space seems less efficient. From Fig. 6(b), variants without "detach" operation generally exceed those with "detach" operation. This can be ascribed to the unprecise supervised signal at the early stage. Among these approaches, OT without "detach" operation achieves the best performance and certifies the effectiveness. If we omit the feature mimicking loss and retrain model, which is displayed as "Non" in Fig. 6(b), PCKh score of this variant reaches 88.09 and the improvement over the E&R submodule mainly comes from the data augmentation. However, it is not the main concern of this work. If we utilize distillation to replace the siamese framework, we can notice that distillation-based approach results in no improvement at all. For visualization, we provide several examples shown in Fig. 1. Predictions under severe occlusions get improved via our approach.

5 Conclusion

The paper proposes an occlusion-aware siamese network for human pose estimation. Firstly, erasing and reconstruction submodule is utilized to erase and reconstruct the occluded features. Secondly, to improve the quality of reconstruction, we propose the siamese framework which enforces the occluded branch to mimic the behavior of the occluded branch. Finally, we employ optimal transport divergence to narrow the distribution discrepancy of these two branches. We conduct our method on three widely used human pose benchmarks and achieve promising results.

Acknowledgements This work was supported by the Research and Development Projects in the Key Areas of Guangdong Province (No.2019B010153001), National Natural Science Foundation of China under Grants 61772527, 61976520 and 61806200. This work was also supported by the Technology Cooperation Project of Application Laboratory, Huawei Technologies Co., Ltd. (FA2018111061-2019SOW05).

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. pp. 3686–3693 (2014)
2. Chen, L., Dai, S., Pu, Y., Zhou, E., Li, C., Su, Q., Chen, C., Carin, L.: Symmetric variational autoencoder and connections to adversarial learning. In: International Conference on Artificial Intelligence and Statistics. pp. 661–669 (2018)
3. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018)
4. Chen, Y., Shen, C., Wei, X.S., Liu, L., Yang, J.: Adversarial posenet: A structure-aware convolutional network for human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1212–1221 (2017)
5. Chou, C.J., Chien, J.T., Chen, H.T.: Self adversarial training for human pose estimation. arXiv preprint arXiv:1707.02439 (2017)
6. Chu, X., Ouyang, W., Li, H., Wang, X.: Structured feature learning for pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4715–4723 (2016)
7. Chu, X., Ouyang, W., Wang, X., et al.: Crf-cnn: Modeling structured information in human pose estimation. In: Advances in Neural Information Processing Systems. pp. 316–324 (2016)
8. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1831–1840 (2017)
9. Genevay, A., Peyré, G., Cuturi, M.: Learning generative models with sinkhorn divergences. arXiv preprint arXiv:1706.00292 (2017)
10. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision. pp. 34–50. Springer (2016)
11. Jain, A., Tompson, J., Andriluka, M., Taylor, G.W., Bregler, C.: Learning human pose estimation features with convolutional networks. arXiv preprint arXiv:1312.7302 (2013)
12. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: bmvc. vol. 2, p. 5. Citeseer (2010)
13. Ke, L., Chang, M.C., Qi, H., Lyu, S.: Multi-scale structure-aware network for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 713–728 (2018)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Li, H., Dai, B., Shi, S., Ouyang, W., Wang, X.: Feature intertwiner for object detection. In: International Conference on Learning Representations (2018)
16. Lin, T.Y., Maire, M., Belongie, S., 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. Springer (2014)
17. Liu, W., Chen, J., Li, C., Qian, C., Chu, X., Hu, X.: A cascaded inception of inception network with attention modulated feature fusion for human pose estimation. In: AAAI (2018)
18. Lu, Y., Chen, L., Saidi, A.: Optimal transport for deep joint transfer learning. arXiv preprint arXiv:1709.02995 (2017)

19. Marras, I., Palasek, P., Patras, I.: Deep globally constrained mrfs for human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3466–3475 (2017)
20. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. Computer vision and image understanding **81**(3), 231–268 (2001)
21. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. pp. 483–499. Springer (2016)
22. Nie, X., Feng, J., Zuo, Y., Yan, S.: Human pose estimation with parsing induced learner. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2100–2108 (2018)
23. Ning, G., Zhang, Z., He, Z.: Knowledge-guided deep fractal neural networks for human pose estimation. IEEE Transactions on Multimedia **20**(5), 1246–1259 (2018)
24. Rafi, U., Leibe, B., Gall, J., Kostrikov, I.: An efficient convolutional network for human pose estimation. In: BMVC. vol. 1, p. 2 (2016)
25. Salimans, T., Zhang, H., Radford, A., Metaxas, D.: Improving gans using optimal transport. arXiv preprint arXiv:1803.05573 (2018)
26. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7912–7921 (2019)
27. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3960–3969 (2017)
28. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5693–5703 (2019)
29. Tang, W., Wu, Y.: Does learning specific features for related parts help human pose estimation? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1107–1116 (2019)
30. Tang, W., Yu, P., Wu, Y.: Deeply learned compositional models for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 190–206 (2018)
31. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning **4**(2), 26–31 (2012)
32. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 648–656 (2015)
33. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in neural information processing systems. pp. 1799–1807 (2014)
34. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1653–1660 (2014)
35. Wang, W., Xu, H., Wang, G., Wang, W., Carin, L.: An optimal transport framework for zero-shot learning. arXiv preprint arXiv:1910.09057 (2019)
36. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4724–4732 (2016)

37. 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. 466–481 (2018)
38. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
39. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 1290–1299. IEEE (2017)
40. Zhang, H., Ouyang, H., Liu, S., Qi, X., Shen, X., Yang, R., Jia, J.: Human pose estimation with spatial contextual information. arXiv preprint arXiv:1901.01760 (2019)
41. Zhou, L., Chen, Y., Wang, J., Lu, H.: Progressive bi-c3d pose grammar for human pose estimation. In: AAAI. pp. 13033–13040 (2020)
42. Zhou, L., Chen, Y., Wang, J., Tang, M., Lu, H.: Bi-directional message passing based scanet for human pose estimation. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). pp. 1048–1053. IEEE (2019)