Journal Pre-proof

PCLoss: Fashion Landmark Estimation with Position Constraint Loss

Meijia Song, Hong Liu, Wei Shi, Xia Li

PII: S0031-3203(21)00215-6
DOI: https://doi.org/10.1016/j.patcog.2021.108028
Reference: PR 108028

To appear in: *Pattern Recognition*

Received date: 26 May 2020
Revised date: 7 November 2020
Accepted date: 1 May 2021

Please cite this article as: Meijia Song, Hong Liu, Wei Shi, Xia Li, PCLoss: Fashion Landmark Estimation with Position Constraint Loss, *Pattern Recognition* (2021), doi: https://doi.org/10.1016/j.patcog.2021.108028

Highlights

- Help network search landmark positions in a more reasonable region
- Alleviate outliers and duplicate detection problems in fashion landmark estimation by loss function
- Easily applied to many popular CNN models without extra computation during inference
- Introduce the skeleton-like characteristic of fashion landmarks to strength the position constraint

PCLoss: Fashion Landmark Estimation with Position Constraint Loss⋆

Meijia Song[a], Hong Liu[a], Wei Shi[a] and Xia Li[a]

[a]Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School

ARTICLE INFO

Keywords:
Fashion landmark estimation
Position constraint
Loss function
Skeleton-like characteristic

ABSTRACT

Fashion landmark estimation aims at locating functional key points of clothes, which has wide potential applications in electronic commerce. However, due to the occlusion and weak outline information, landmark estimation occurs outliers and duplicate detection problems. To alleviate these issues, we propose Position Constraint Loss (PCLoss) to constrain error landmark locations by utilizing the position relationship of landmarks. Specifically, PCLoss adds a regularization term for each landmark to regularize their relative positions, and it can be easily applied to both regression and heatmap based methods without extra computation during inference. Unlike existing approaches that propagate landmark information between feature layers by specific network structures, PCLoss introduces position relations of landmarks in the label space without modifying the network structure. In addition, we leverage the skeleton-like relation of clothing to further strengthen position constraints between landmarks. Extensive experimental results on DeepFashion, FLD and FashionAI demonstrate that our methods can effectively increase the performance of mainstream frameworks by a large margin. We also explore the effectiveness of PCLoss on human pose estimation task, and the experimental results on COCO 2017 prove the generality of our methods on other key point estimation tasks.

1. Introduction

With the rapid development of online commerce, fashion image analysis, such as clothing retrieval[8, 9, 12, 15, 28], classification[1, 2, 20, 25, 29, 32] and matching recommendation[13, 14, 25], shows great potential in online industry. However, large deformation and appearance changes in different clothing categories, scales and shapes make fashion image analysis difficult. To capture shapes and structures of clothes, a more discriminative representation, fashion landmark, has been proposed to support these high-level applications. Similar to human joints, fashion landmarks are functional key points defined on clothes, such as the cuff, neckline and waistline. As clothes are non-rigid and have large deformation, the fashion landmark estimation still remains challenging.

Even with big training data[6, 7, 18, 19, 20, 21, 34] and advanced network architectures, fashion landmark estimation still has several problems: (1) *Outliers*: the estimator is confused by the high response area of background or occlusion, resulting in some outliers in the predicted result as shown in Figure 1(a). (2) *Duplicate detection*: the estimator repeatedly detects a certain key point due to large deformation or weak outline information in clothing as shown in Figure 1(b).

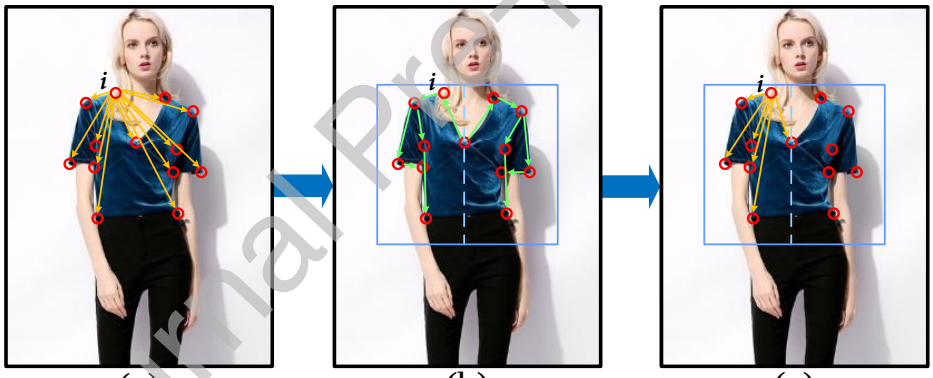
There were some researches[3, 5, 29, 33] that attempted to alleviate above problems by leveraging position relationships between landmarks. Cao et al.[3] proposed the concept of Part Affinity Fields (PAFs) to measure the association of body parts, and used different branches of network to learn heatmaps and PAFs respectively. Chu et al.[5] designed geometrical transform kernels for the network to capture landmark relations in different feature channels. Wang et al.[29] designed two fashion grammars to describe landmark relations, and proposed a Bidirectional Convolutional Recurrent Neural Network (BCRNN) to perform the message passing over the fashion grammars. Yu et al.[33] proposed layout-graph reasoning layers to introduce structural landmark relationships to the network. These methods unanimously focused on the optimization of network structures and tried to learn the relations between landmark positions at feature level. However, such practices relied on specific network structures, which would limit the generalization of models and increase the inference complexity.

⋆This work is supported by National Key RD Program of China (No. 2020AAA0108904), Basic and Applied Basic Research Foundation of Guangdong (No.2020A1515110370), Science and Technology Plan of Shenzhen (Nos.JCYJ20190808182209321).

songmj@pku.edu.cn (M. Song); hongliu@pku.edu.cn (H. Liu); pkusw@pku.edu.cn (W. Shi); ethanlee@pku.edu.cn (X. Li)
ORCID(s):

(a) Outliers (b) Duplicate detection

Figure 1: Visualization of outliers and duplicate detection problems in fashion landmark estimation. In these images, the predicted landmarks are indicated by blue dots and the ground truth landmarks by orange dots. In (a), the 3-th point (labeled by red dotted rectangle) is the outlier among these landmarks. In (b), the 7-th and 8-th points (labeled by red dotted rectangle) stand for the duplicate detection of the cuff. (Best viewed in color)



(a) (b) (c)

Figure 2: Visualization of skeleton-like characteristic. (a) The position relationship between landmark i and others. (b) The skeleton-like structure of fashion landmarks. According to the skeleton-like characteristic, the landmarks can be divided into two parts as the blue rectangle in the image. After the optimization, the position relationship between landmark i and other points can be described as (c).

In this paper, we propose a more efficient solution, Position Constraint Loss (PCLoss), which is used to regularize relative positions of landmarks during training only. Instead of designing a specific network structure to capture landmark relations, our method adds a regularization term for each landmark by loss function to correct error points, which can be easily applied to most popular models without modifying network structures. Previous works mainly supervise prediction of each key point with $L2$ or MSE loss directly. PCLoss, instead, constrains relative positions between points. Concretely, when there exist some landmarks whose relative positions with others are far from normal values, more penalties will be imposed on them to regularize their locations. With PCLoss, the network will learn to search for landmark locations in a more reasonable region.

Furthermore, to reduce the influence of pose variation and error points on PCLoss, the skeleton-like optimization is introduced to the loss function. Since each landmark in PCLoss is associated with all remaining points, once there are some error points, all landmarks will be affected due to position constraints. To alleviate the situation, we propose the

skeleton-like structural constrain mechanism to associate the target landmark with only high-related points. According to skeleton-like features, we can divide fashion landmarks into two parts as shown in Figure 2. In each part, landmarks satisfy the skeleton-like relation and possess stronger correlations with each other. With skeleton-like optimization, PCLoss will have more effective position constraints for landmarks.

The contributions of this paper can be summarized as follows: (1) We propose a position constraint loss to alleviate outliers and duplicate detection problems in fashion landmark estimation, and the loss can be easily applied to many popular CNN models without extra computation during inference. (2) Inspired by the knowledge of human body structure, we introduce the skeleton-like characteristic of fashion landmarks to further optimize the position constraint loss.

To demonstrate the effectiveness of our PCLoss, we apply the loss on some advanced base models, such as SimpleBaselines [30], FPN [16]. What's more, we evaluate the performance of PCLoss on both regression and heatmap based methods. Extensive experiments on three large-scale fashion datasets, namely DeepFahsion, FLD and FashionAI, demonstrate that, our PCLoss can effectively increase the performance of mainstream frameworks by a large margin. To further explore the performance of PCLoss on other key point estimation tasks, we extend it to human pose estimation task. The experimental results on COCO[17] 2017 proves that our method has good generality on key point estimation tasks.

2. Related Work

Fashion landmark estimation: In recent years, fashion image analysis has attracted more and more attention. In particular, some large-scale fashion datasets[18, 19, 34] such as DeepFashion[20], DeepFashion2[6], FLD[21] and FashionAI[7] have been proposed, which further promote the development of fashion image analysis. Fashion image analysis includes several tasks, such as clothing detection[4, 6, 34], retrieval[8, 9, 12, 15, 28], classification[1, 2, 25, 32], landmark estimation[21, 29, 31] and so on. Among these tasks, fashion landmark estimation is a more fine-grained work, which focuses on the shape and structure of clothing. The methods of fashion landmark estimation can be divided into two categories: regression based methods and heatmap (a confidence map of positional distribution for each landmark) based methods. Models based on regression methods are differentiable, which can directly regress the coordinate of each landmark. Liu et al.[21] used a cascade network to regress landmark positions, and the coordinates would be refined by each stage of the network. Yan et al.[31] designed a spatial transformer network based on regression methods to solve the environmental influence on landmark estimation. While in recent studies, researchers preferred the heatmap[22, 27] based methods which are able to capture more spatial information than regression based methods. Wang et al.[29] used an attention model to output the heatmap results for landmarks. Huang et al.[11] designed a Match R-CNN (based on Mask R-CNN[10]) to estimate heatmaps. As our position constraint loss utilizes the coordinate information to calculate relative positions for landmarks, it can be easily applied to regression based methods. To suit heatmap based methods, we use the integral operation[26] to calculate the landmark coordinates first, then apply our position constraint loss.

Key points correlation models: Many researches attempted to utilize the position relationship between key points to help the network learn more reasonable features. Cao et al.[3] proposed a two-branch network to study heatmaps and Part Affinity Fields (describe the correlation of different body parts), respectively. As Chu et al.[5] found that body joints had their own feature channels, they designed a bi-directional tree to propagate the information between them. Similar with this, Wang et al.[29] introduced two fashion grammars to describe landmark relations and embedded BCRNN units into the CNN model to simulate the message passing over fashion grammars. Yu et al.[33] designed layout-graph reasoning layers to leveraging the landmark relationship for structural results. All these researches tried to utilize specific network structures to capture the position relationship between key points. However, such solutions will bring more inference overhead. In this paper, we attempt to solve the problem from a new perspective, introducing the positional correlation of key points by using a simple position constraint loss. Moreover, inspired by the researches [5, 29], which passed the messages over feature maps according to a specific structure, we apply the skeleton-like characteristic in PCLoss to further enhance the position constraint of the loss function.

3. Position Constraint Loss

In this paper, we evaluate the effectiveness of PCLoss in two kind of methods: regression based methods and heatmap based methods. The overall pipeline of our framework is illustrated in Figure 3. On both regression and

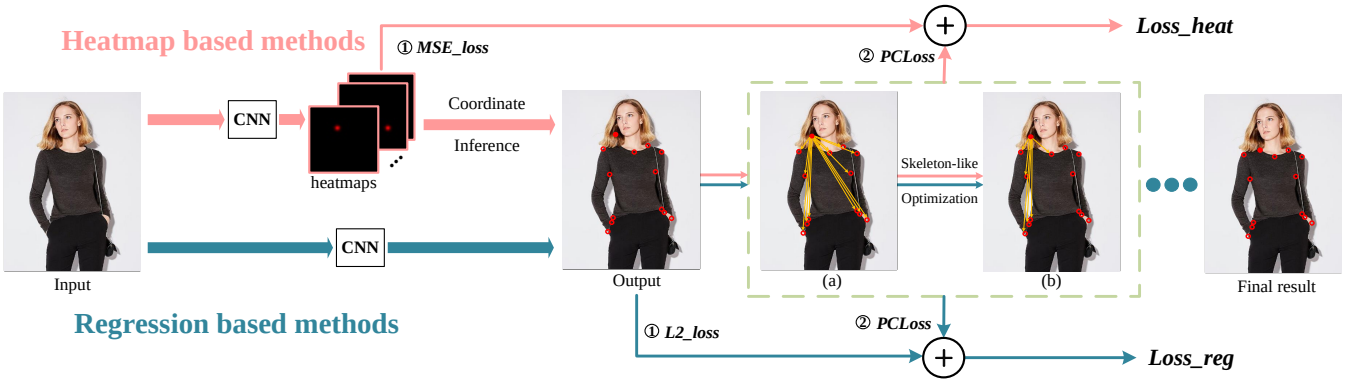


Figure 3: Pipeline of our method. The upper branch (denoted in pink) shows the pipeline of heatmap based methods and the bottom branch (denoted in blue) is the pipeline of regression based methods. In regression based methods, we can obtain the landmark coordinates directly. While in heatmap based methods, the network will output the heatmap result for each landmark and we need to infer their coordinates by integral operation. (a) The original position correlations between landmarks in PCLoss. (b) The position correlations after skeleton-like optimization. (Best viewed in color)

heatmap based models, landmark coordinates should be acquired first. In regression based models, landmark coordinates can be obtained from outputs directly. While in heatmap based models, we need to infer landmark coordinates from heatmap results first. Then, the coordinates can be used to calculate PCLoss. After that, the skeleton-like relation is employed to strengthen the position constraints between landmarks.

In this section, we will introduce our method from four parts. Firstly, we will describe the coordinate inference method for heatmap based models. Then, the definition of PCLoss will be given in detail. After that, a skeleton-like based optimization method will be introduced to PCLoss. Finally, the overall loss function will be described for both regression and heatmap based methods.

3.1. Coordinate Inference for Heatmaps

In regression based methods, we can easily obtain landmark locations and train position coordinates end-to-end, while in heatmap based methods, networks output the heatmap result for each landmark instead. Therefore, in heatmap based models, landmark coordinates should be inferred from heatmaps first for the calculation of PCLoss.

Mainstream methods use maximum likelihood[26] to infer landmark coordinates from heatmaps as

$$p_i = \arg\max_l H_i(l) \tag{1}$$

where H_i represents the heatmap for landmark i. l denotes the location (l_x, l_y) in the heatmap, and p_i is the coordinate (x_i, y_i) of the landmark i inferred from the location l with the maximum likelihood.

Although this method can obtain accurate results from heatmaps, it is not differentiable. Therefore, with maximum likelihood, PCLoss can not be trained end-to-end in heatmap based models. To solve the problem, we adopt the integral operation[26] to calculate landmark coordinates as

$$\tilde{H}_i(l) = \frac{e^{\alpha \cdot H_i(l)}}{\sum_l e^{\alpha \cdot H_i(l)}} \tag{2}$$

$$p_i = \sum_{l_y=1}^{H} \sum_{l_x=1}^{W} l \cdot \tilde{H}_i(l) \tag{3}$$

where $\tilde{H}_i(l)$ is the heatmap after normalization. The size of heatmap is $H \times W$. α is the scale factor and set as 0.1 in this paper. With the integral operation, landmark coordinates are trainable in heatmap based methods.

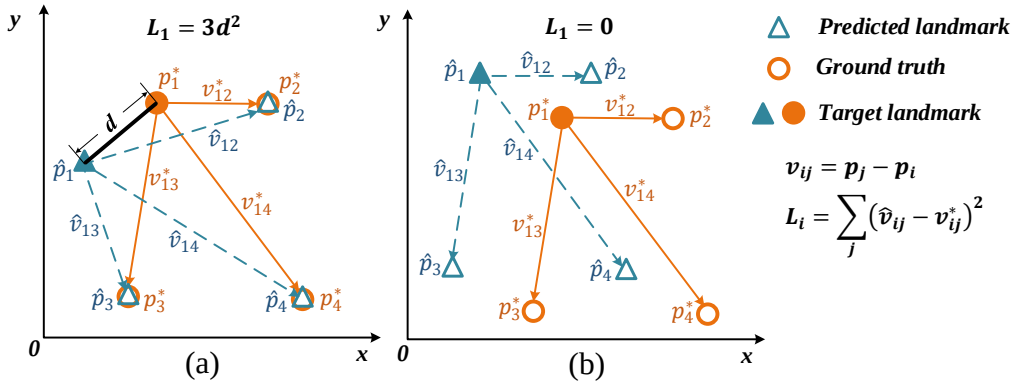


Figure 4: Examples of position constraint loss. p_i denotes the landmark (x_i, y_i) and $\vec{v}_{ij}$ denotes the relative position vector between landmark p_i and p_j.

3.2. Definition of Position Constraint Loss

Position Constraint Loss is designed for regularizing relative positions between landmarks. Assume that there are N landmarks defined on clothes, the goal of fashion landmark estimation is to predict the position P_t for all landmarks as follow

$$P_t = \{p_i, i = 1, 2, ..., N\} \tag{4}$$

As PCLoss pays attention to the position relationship of two landmarks, the relative position vector between landmark i and j is given by

$$v_{ij} = p_j - p_i \tag{5}$$

where $i, j \in [1, N]$. Then, PCLoss for the landmark i can be defined as $L2$ loss between predicted and ground truth relative position vector, which is formulated as

$$L_i = \sum_{j=1}^{N} (\hat{v}_{ij} - v_{ij}^*)^2 \tag{6}$$

where $\hat{v}_{ij}$ and v_{ij}^* are the predicted and ground truth relative position vector. L_i is the PCLoss for the landmark i. Therefore, PCLoss for all landmarks L_{lan} can be described as

$$L_{lan} = \{L_1, L_2, \cdots, L_N\} \tag{7}$$

Figure 4 presents two examples of PCLoss. In Figure 4(a), all predicted landmark positions are correct except $\hat{p}_1$, which leads to a large PCLoss. While in Figure 4(b), all predicted landmarks are panned but PCLoss is equal to zero. From these examples, we can easily find that PCLoss actually regularizes relative positions between landmarks instead of absolute positions. However, this solution will bring a question: although some landmarks are accurate enough that no additional penalties are required, their position constraint losses still have large values due to outliers. To alleviate the problem, we adopt the concept of OHEM[24] algorithm which improves the network performance by dealing with hard examples. In this paper, we only calculate k max PCLoss in L_{lan} like OHEM. Then, the final PCLoss L_{PC} can be formulated as

$$L_{PC} = f_k\{L_{lan}\} \tag{8}$$

where $f_k\{\cdot\}$ is a function that calculates the average of k maxima of the set.

Through PCLoss, the position relationship of landmarks will be taken into account by the network. As a result, outliers and duplicate detection problems will be mitigated and the network will learn to search for landmark locations in a more reasonable region.

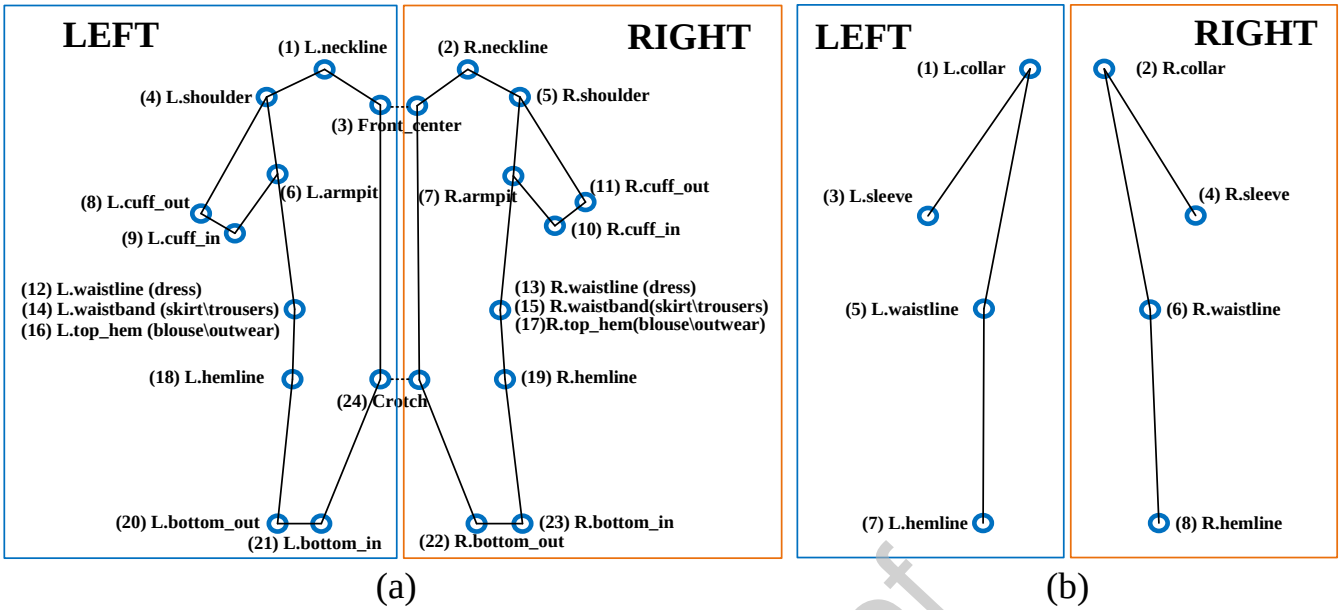


Figure 5: Visualization of skeleton-like structure. (a) The skeleton-like structure based on the 24 landmark model. (b) The skeleton-like structure based on the 8 landmark model. According to the skeleton-like characteristic introduced in §3.2, the landmarks can be divided into two parts. In each part, landmarks satisfy the skeleton-like constraint. Specifically, on the 24 landmark model, the middle landmarks like front center and crotch will be shared by both parts as shown in (a).

3.3. Optimization Based on Skeleton-Like Relation

Fashion landmarks are similar to human joints, satisfying the skeleton-like and symmetry relations. Since the symmetry characteristic is the relation between landmark pairs, it is not suitable for the PCLoss. Therefore, we introduce the skeleton-like characteristic to further optimize the loss design.

Based on prior knowledge, landmark positions will have large variances when the human pose changes. If we connect a key point with all landmarks defined on clothes, PCLoss will be too sensitive to the pose variation. To optimize the position constraints, for a certain landmark, we filter the unimportant points and just connect it with more relavant landmarks.

As Figure 5, we divide fashion landmarks into two parts and the landmarks from each part satisfy the skeleton-like relation. After the optimization, each landmark will have a stronger position correlation with others. According to the skeleton-like relation, we divide fashion landmarks into two parts as follows

$$\begin{aligned} P_l &= \{p_{l_1}, p_{l_2}, \dots, p_{l_M}\} \\ P_r &= \{p_{r_1}, p_{r_2}, \dots, p_{r_M}\} \end{aligned} \tag{9}$$

where P_l and P_r are landmark positions in left part and right part, and each part contains M landmarks. The partition rule is illustrated in Figure 5. Notably, in the structure of 24 landmark models, the middle points, like front center and crotch, will be shared by two sets.

Then, we reuse Eq. (5)– (7) to calculate PCLoss L^l_{lan} and L^r_{lan} for P_l and P_r, respectively. After that, the final PCLoss after optimization can be formulated as

$$L_{PC} = f_k\{L^l_{lan}\} + f_k\{L^r_{lan}\} \tag{10}$$

With skeleton-like optimization, each landmark is only associated with others that satisfy the skeleton-like relation and the position constraints between landmarks will become stronger. In addition, we have also considered other strategies to divide landmarks. For example, we have ever divided landmarks into top and bottom part, but it can only achieve few improvements due to the imbalanced division mode (e.g. when the clothes type is blouse, it only contains top landmarks and the bottom part is useless). But with skeleton-like strategy, landmarks will be divided into left and right part, and there will always be valid landmarks in each part in spite of clothes types.

Figure 6: Visualization of fashion landmark estimation. Our results are labeled by red dotted rectangles. (Best viewed in color)

3.4. Overall Loss Function

In this paper, we evaluate the effectiveness of PCLoss in two kind of models: regression based models and heatmap based models. In regression based models, the $L2$ loss and PCLoss are both used to train the network as

$$L_{reg} = \alpha_r \cdot L2(\hat{p}, p^*) + \beta_r \cdot L_{PC}(\hat{p}, p^*) \tag{11}$$

where the $L2$ loss is used to measure the distance between the predicted landmark coordinate $\hat{p}$ and the ground truth landmark coordinate p^*, and the L_{PC} is the position constraint loss proposed in this paper. α_r and β_r are the balancing weights of the two losses.

Similar to the L_{reg}, heatmap based methods adopt MSE and PCLoss to train the network as

$$L_{heat} = \alpha_h \cdot MSE(\hat{h}, h^*) + \beta_h \cdot L_{PC}(\hat{p}, p^*) \tag{12}$$

where the MSE loss is applied to calculate the error between the predicted heatmap $\hat{h}$ and the ground truth heatmap h^*. α_h and β_h are the balancing weights in L_{heat}.

Notably, since PCLoss is larger than $L2$ loss and MSE loss, the hyper parameters β_r and β_h should be smaller than α_r and α_h to keep balance.

Table 1
Comparison results on the DeepFashion dataset with Normalized Error

Model	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waist	R.Waist	L.Hem	R.Hem	NE (avg.)
FashionNet[20](reg)	8.54%	9.02%	9.73%	9.35%	8.54%	8.45%	8.12%	8.23%	8.75%
DFA[21](reg)	6.28%	6.37%	6.58%	6.21%	7.26%	7.02%	6.58%	6.63%	6.62%
DLAN[31](reg)	5.70%	6.11%	6.72%	6.47%	7.03%	6.94%	6.24%	6.27%	6.44%
Ours+FashionNet	**2.33%**	**2.40%**	**4.00%**	**4.17%**	**5.81%**	**6.01%**	**2.13%**	**2.20%**	**3.63%**
FGN[29](heat)	4.15%	4.04%	4.96%	4.49%	5.02%	5.23%	5.37%	5.51%	4.85%
FPN[16](heat)	2.19%	2.19%	3.28%	3.31%	4.81%	4.87%	2.09%	2.11%	3.11%
SPB[30](heat)	2.07%	2.10%	3.18%	3.13%	4.83%	4.83%	1.75%	1.77%	2.96%
Ours+FPN	**2.04%**	**2.05%**	**3.18%**	**3.19%**	**4.72%**	**4.81%**	**1.90%**	**1.96%**	**2.98%**
Ours+SPB	**2.03%**	**2.06%**	**3.00%**	**3.04%**	**4.64%**	**4.75%**	**1.50%**	**1.53%**	**2.82%**

Our methods are marked in bold and lower values are better. The label 'reg' means the model is based on regression methods, and 'heat' means the model is based on heatmap methods.

Table 2
Comparison results on the FLD dataset with Normalized Error

Model	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waist	R.Waist	L.Hem	R.Hem	NE (avg.)
FashionNet[20](reg)	7.84%	8.03%	9.75%	9.23%	8.74%	8.21%	8.02%	8.93%	8.59%
DFA[21](reg)	4.80%	4.80%	9.10%	8.90%	–	–	7.10%	7.20%	6.98%
DLAN[31](reg)	5.31%	5.47%	7.05%	7.35%	7.52%	7.48%	6.93%	6.75%	6.73%
Ours+FashionNet	**3.86%**	**3.94%**	**7.46%**	**7.38%**	**7.70%**	**7.65%**	**5.00%**	**4.95%**	**5.99%**
FGN[29](heat)	4.63%	4.71%	6.27%	6.14%	6.35%	6.92%	6.35%	5.27%	5.83%
FPN[16](heat)	2.83%	2.89%	5.19%	5.22%	6.60%	6.57%	4.73%	4.50%	4.82%
SPB[30](heat)	2.88%	2.89%	5.11%	5.20%	6.53%	6.28%	4.48%	4.37%	4.72%
Ours+FPN	**2.83%**	**2.84%**	**5.21%**	**5.17%**	**6.56%**	**6.49%**	**4.31%**	**4.30%**	**4.71%**
Ours+SPB	**2.86%**	**2.84%**	**5.01%**	**5.05%**	**6.44%**	**6.28%**	**4.18%**	**3.95%**	**4.58%**

Our methods are marked in bold and lower values are better. The label 'reg' means the model is based on regression methods, and 'heat' means the model is based on heatmap methods. '–' denotes the detailed results which are not released.

4. Experiments and Analysis

In this section, we evaluate the performance of our PCLoss on three large datasets, DeepFashion, FLD and FashionAI. Firstly, we will introduce the information of datasets and describe the details of implementation. Then, the experimental results and analysis based on regression and heatmap methods will be given. After that, extensive ablation studies are performed to demonstrate the effectiveness of different modules. Besides, We extend PCLoss to human pose estimation task.

4.1. Datasets

DeepFashion[20] contains three subsets for different benchmarks and we choose the largest one (subset for Category and Attribute Prediction) to evaluate the performance of fashion landmark estimation. The DeepFashion is the largest one among the three datasets. It contains 289,222 images, which are taken under various scenarios. Each image from DeepFashion has rich annotations, including category, bounding box, visibility and 8 landmarks. Following the dataset setting, the training, validation and testing set have 209,222, 40,000 and 40,000 images, respectively.

FLD[21] is a dataset specifically designed for fashion landmark estimation, which contains 123,016 clothes images. Similar to DeepFashion, each image of FLD is also annotated with category, bounding box, visibility and 8 landmarks. Following dataset setting, there are 83,033, 19,992 and 19,992 images for training, validation and testing.

FashionAI[7] is a competition dataset proposed by Alibaba Group. It has rich landmark annotations that each image

Table 3
Comparison results on FashionAI-val dataset with NE

Model	NE (avg.)
FashionNet[20](reg)	10.17%
Ours+FashionNet[20]	**9.71%**
FPN[16](heat)	4.12%
SPB[30](heat)	4.10%
Ours+FPN	**3.91%**
Ours+SPB	**3.90%**

Our methods are marked in bold and lower values are better. The label 'reg' means the model is based on regression methods, and 'heat' means the model is based on heatmap methods.

is annotated with category, visibility and 24 landmarks. The FashionAI dataset covers 5 women clothing categories and has 104,088 clothes images. Since there is not validation set in FashionAI, we build the validation set by selecting the previous 2,100, 1,900, 1,900, 2,100 and 2,000 images from the blouse, outwear, dress, skirt and trousers category in training set. After that, there are 66,866, 10,000 and 27,222 images in training, validation and testing set, respectively.

4.2. Experiment Settings

Evaluation Metrics. Following the common practice[7, 21, 29], we adopt Normalized Error (NE)[21] as our evaluation metrics. NE is the average normalized distance between predicted landmark position and ground truth position as formulated as

$$NE = \frac{\sum_{i=0}^{N} \frac{1}{T_i} \sqrt{(\hat{p}_i - p_i{}^*)^2} \cdot v_i}{\sum_{i=0}^{N} v_i} \times 100\% \tag{13}$$

where v_i is the visibility of landmark i and T_i is the normalization coefficient.

On DeepFashion and FLD, the coefficient T is the width of the image. But on FashionAI, T is the Euclidean distance between a specific pair of landmarks defined on FashionAI (for blouse, outwear and dress, they are two armpit points; and for trousers and skirt, they are two waistband points).

Training Details. Our experiments are implemented based on Pytorch framework. We use Adam optimizer over 4 GPUs with a total of 20 images per minibatch (5 images per GPU). We train the network for $100k$ and $150k$ iterations with an initial learning rate of 0.0003, which is divided by 3 at 30k and again at 50k iterations. Before training, the images from DeepFashion and FLD are cropped by bounding box and all images from three datasets are resized to 512×512. Data augmentation such as scale (±25%), rotation (±30°) and flip, are also applied in training. We calculate the top 8 PCLoss on FashionAI, and top 4 PCLoss on DeepFashion and FLD dataset.

4.3. Performance Evaluation

DeepFashion and FLD dataset. To demonstrate the effectiveness of our PCLoss, we apply it to two kind of models: regression based models and heatmap based models. As shown in Table 1 and Table 2, FashionNet[20], DFA[21] and DLAN[31] are all the regression based models, while FGN[29], FPN[16] and SPB(Simplebaselines)[30] are the heatmap based models. To test the performance of PCLoss on regression based methods, we apply our PCLoss to FashionNet model, which is the baseline proposed in DeepFashion dataset. Besides, we train the FPN and SPB network by our PCLoss to examine the effectiveness of the loss on heatmap based methods. The comparison results on DeepFashion and FLD dataset are presented in Table 1 and Table 2. From the table, we can easily find that, on both model types (regression based and heatmap based models), our position constraint loss can effectively improve the location accuracies of networks.

FashionAI dataset. Since FashionAI is a competition dataset, the ground truth annotations from testing set are not released. Therefore, we use the validation set divided in §4.1 to better compare the performance between different models. On FashionAI dataset, we apply our PCLoss to three deep models, including FashionNet[20], FPN[16] and

Table 4
Ablation studies on FashionAI validation dataset

Model	PCLoss	Skeleton-like	Top-bottom	NE (avg.)	△
FashionNet	-	-	-	10.1724%	0.4646%
FashionNet	✓	-	-	9.8295%	0.1217%
FashionNet	✓	✓	-	**9.7078%**	-
FPN	-	-	-	4.1219%	0.2109%
FPN	✓	-	-	3.9646%	0.0536%
FPN	✓	-	✓	3.9597%	0.0487%
FPN	✓	✓	-	**3.9110%**	-

△ denotes the difference value between our methods (marked in bold) and the others.

SPB(Simplebaselines)[30], as illustrated in Table 3. The FashionAI is the most challenging one among the three datasets due to more landmark annotations and lack of bounding boxes. As FashionAI dataset has 24 landmarks to predict, it is more likely to encounter the outliers and duplicate detection problems. Therefore, the position constraints between landmarks can play a greater role on FashionAI. Compared with the results on DeepFashion and FLD dataset, PCLoss can lead to greater performance gains on FashionAI. However, as regression based methods are hard to deal with the more landmark estimation problem due to lack of spatial information, the result of FashionNet on FashionAI is worse than DeepFashion and FLD. In Figure 6, we present some picture results on FashionAI. From results, we can find that PCLoss can help the network search landmark positions in a more reasonable region, and effectively mitigate the influence of occlusion and weak outline information.

4.4. Ablation Study

To demonstrate the merits of each component in PCLoss, ablation studies are performed in this subsection. The experimental results are summarized in Table 4.

Effectiveness of position constraint loss and skeleton-like optimization. To prove the effectiveness of our PCLoss and skeleton-like optimization method, we perform the experiments on the validation set of FashionAI and the results are presented in Table 4. We evaluate our method on both regression based model (FashionNet) and heatmap based model (FPN). From the results, we can find that even the original PCLoss (without skeleton-like optimization) is able to promote the network performance by 0.3429% (FashionNet) and 0.1573% (FPN). Then we apply the skeleton-like optimization to the loss, and the accuracies can be further promoted by 0.1217% (FashionNet) and 0.0536% (FPN).

PCLoss optimization with different division modes. Skeleton-like strategy is an optimization method to help PCLoss obtain better results without extra annotations. Besides skeleton-like strategy, there are still have other division modes that can strengthen the constraint of PCLoss. In this experiment, we compare the top-bottom strategy with the skeleton-like strategy to explore the influence of different division modes. In the top-bottom strategy, we divide landmarks into top and bottom parts. As shown in Table 4, top-bottom strategy can bring few improvements for PCLoss due to the imbalanced division mode. Specifically, as Figure 7 shows, when the clothes type is blouse, the top part includes all visible landmarks and the bottom part is useless. But with skeleton-like strategy, landmarks will be divided into left and right parts, and both parts will have valid landmarks in spite of clothes types. Therefore, skeleton-like strategy is a more balanced division mode compared to other strategies.

Selection of loss weights. As Eq. (11) and Eq. (12), we use the multi-loss to train the network. In the paper, we set the $\alpha_r = \alpha_h = \alpha = 1$ and $\beta_r = \beta_h = \beta$. As the value of position constaint loss is larger than $L2$ and MSE loss, the L_{PC} loss weight β should be smaller than α to keep balance. In Figure 8, we show the network performance in different values of β. According to the experiments, we set the $\beta = 2 \times 10^{-4}$ on FashionAI dataset, and $\beta = 1 \times 10^{-5}$ on DeepFashion and FLD dataset.

4.5. Timing

Since PCLoss improves the network performance through loss function, it will not bring extra time consumption during inference. PCLoss is also fast to train. The calculation of PCLoss using 1-GPU (20 images per mini-batch) on FashionAI only takes 2.433 milliseconds. Therefore, PCLoss is an efficient way to increase the locating accuracy for key point estimation tasks.

Figure 7: Comparison between different division modes.

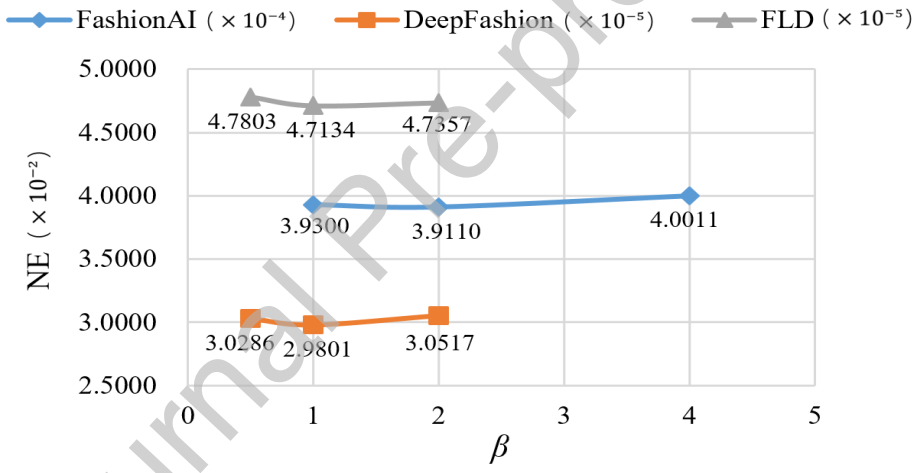


Figure 8: Comparison of the loss weight β on the DeepFashion, FLD and FashionAI dataset.

4.6. PCLoss for Human Pose Estimation

PCLoss is designed to help the network regularize key point positions, therefore, it can easily be extended to other key point estimation tasks. In this section, we apply PCLoss to human pose estimation. We use FPN[16] and SPB(Simplebaselines)[30] as base models to examine the effectiveness of PCLoss. All models are trained on the COCO train2017 dataset and tested on the COCO val2017 dataset. For detection, we use a faster-RCNN[23] detector with detection AP 56.4 for the person category on COCO val2017 as [30]. As presented in Table 5, PCLoss can effectively improve the performance of human pose estimation task, which proves the generality of our methods.

5. CONCLUSIONS

In this paper, we design a Position Constraint Loss (PCLoss) for fashion landmark estimation, which incorporates the position correlation into landmark estimation models. Specifically, the PCLoss adds a regular term for each landmark to regularize their relative positions. Compared with other alternatives, our PCLoss effectively mitigates the outliers and duplicate detection problems without modifying existing CNN architectures. In addition, our skeleton-

Table 5
Comparison results on COCO val2017 dataset

Model	Input Size	AP
FPN[16]	256 x 192	69.9
SPB[30]	256 x 192	70.4
Ours+FPN	**256 x 192**	**70.6**
Ours+SPB	**256 x 192**	**70.7**

Our methods are marked in bold and higher values are better.

like optimization method further strengthens the position constraints between landmarks. The proposed method can be applied to both regression and heatmap based methods and it provides a novel perspective towards position relation learning in key point estimation tasks. Extensive experimental results on three challenging datasets, DeepFashion, FLD and FashionAI, demonstrate that our method outperforms other state-of-the-art methods. The experiment on COCO 2017 shows the potential applications of PCLoss for other key point estimation tasks, which can be explored more in future work.

References

[1] Al-Halah, Z., Stiefelhagen, R., Grauman, K., 2017. Fashion forward: Forecasting visual style in fashion, in: Proc. ICCV, pp. 388–397.
[2] Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., Van Gool, L., 2012. Apparel classification with style, in: Proc. ACCV, pp. 321–335.
[3] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y., 2017. Realtime multi-person 2D pose estimation using part affinity fields, in: Proc. CVPR, pp. 7291–7299.
[4] Chen, Q., Huang, J., Feris, R., Brown, L.M., Dong, J., Yan, S., 2015. Deep domain adaptation for describing people based on fine-grained clothing attributes, in: Proc. CVPR, pp. 5315–5324.
[5] Chu, X., Ouyang, W., Li, H., Wang, X., 2016. Structured feature learning for pose estimation, in: Proc. CVPR, pp. 4715–4723.
[6] Ge, Y., Zhang, R., Wu, L., Wang, X., Tang, X., Luo, P., 2019. DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images, in: Proc. CVPR, pp. 5337–5345.
[7] Group, A., . FashionAI dataset http://fashionai.alibaba.com/datasets/.
[8] Hadi Kiapour, M., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L., 2015. Where to buy it: Matching street clothing photos in online shops, in: Proc. ICCV, pp. 3343–3351.
[9] Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S., 2017. Automatic spatially-aware fashion concept discovery, in: Proc. ICCV, pp. 1463–1471.
[10] He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proc. ICCV, pp. 2961–2969.
[11] Huang, C., Chen, J., Pan, Y., Lai, H., Yin, J., Huang, Q., 2018. Clothing landmark detection using deep networks with prior of key point associations, pp. 1–11.
[12] Huang, J., Feris, R.S., Chen, Q., Yan, S., 2015. Cross-domain image retrieval with a dual attribute-aware ranking network, in: Proc. ICCV, pp. 1062–1070.
[13] Kiapour, M.H., Yamaguchi, K., Berg, A.C., Berg, T.L., 2014. Hipster wars: Discovering elements of fashion styles, in: Proc. ECCV, pp. 472–488.
[14] Lee, H., Seol, J., Lee, S.g., 2017. Style2vec: Representation learning for fashion items from style sets. arXiv preprint arXiv:1708.04014 .
[15] Liao, L., He, X., Zhao, B., Ngo, C.W., Chua, T.S., 2018. Interpretable multimodal retrieval for fashion products, in: Proc. ACM MM, pp. 1571–1579.
[16] Lin, T., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J., 2017. Feature pyramid networks for object detection, in: Proc. CVPR, pp. 2117–2125.
[17] Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: Proc. ECCV, pp. 740–755.
[18] Liu, K., Chen, T., Chen, C., 2016a. MVC: A dataset for view-invariant clothing retrieval and attribute prediction, in: Proc. ACM MM, pp. 313–316.
[19] Liu, S., Liang, X., Liu, L., Lu, K., Lin, L., Cao, X., Yan, S., 2015. Fashion parsing with video context. IEEE Transactions on Multimedia , 1347–1358.
[20] Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X., 2016b. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations, in: Proc. CVPR, pp. 1096–1104.
[21] Liu, Z., Yan, S., Luo, P., Wang, X., Tang, X., 2016c. Fashion landmark detection in the wild, in: Proc. ECCV, pp. 229–245.
[22] Pfister, T., Charles, J., Zisserman, A., 2015. Flowing convnets for human pose estimation in videos, in: Proc. ICCV, pp. 1913–1921.
[23] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Proc. NeurIPS, pp. 91–99.

[24] Shrivastava, A., Gupta, A., Girshick, R., 2016. Training region-based object detectors with online hard example mining, in: Proc. CVPR, pp. 761–769.

[25] Simo-Serra, E., Fidler, S., Moreno-Noguer, F., Urtasun, R., 2015. Neuroaesthetics in fashion: Modeling the perception of fashionability, in: Proc. CVPR, pp. 869–877.

[26] Sun, X., Xiao, B., Liang, S., Wei, Y., 2018. Integral human pose regression, in: Proc. ECCV, pp. 529–545.

[27] Tompson, J.J., Jain, A., LeCun, Y., Bregler, C., 2014. Joint training of a convolutional network and a graphical model for human pose estimation, in: Proc. NeurIPS, pp. 1799–1807.

[28] Veit, A., Kovacs, B., Bell, S., McAuley, J., Bala, K., Belongie, S., 2015. Learning visual clothing style with heterogeneous dyadic co-occurrences, in: Proc. ICCV, pp. 4642–4650.

[29] Wang, W., Xu, Y., Shen, J., Zhu, S., 2018. Attentive fashion grammar network for fashion landmark detection and clothing category classification, in: Proc. CVPR, pp. 4271–4280.

[30] Xiao, B., Wu, H., Wei, Y., 2018. Simple baselines for human pose estimation and tracking, in: Proc. ECCV, pp. 466–481.

[31] Yan, S., Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X., 2017. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks, in: Proc. ACM MM, pp. 172–180.

[32] Yang, W., Luo, P., Lin, L., 2014. Clothing co-parsing by joint image segmentation and labeling, in: Proc. CVPR, pp. 3182–3189.

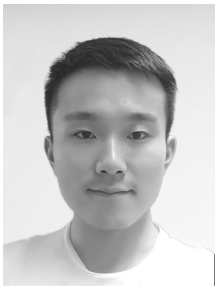
[33] Yu, W., Liang, X., Gong, K., Jiang, C., Xiao, N., Lin, L., 2019. Layout-graph reasoning for fashion landmark detection, in: Proc. CVPR, pp. 2937–2945.

[34] Zheng, S., Yang, F., Kiapour, M.H., Piramuthu, R., 2018. Modanet: A large-scale street fashion dataset with polygon annotations, in: Proc. ACM MM, pp. 1670–1678.

Meijia Song received the bachelor's degree from Chongqing University in 2017. She is currently pursuing her master's degree at Peking University (PKU), China, advised by Prof. Hong Liu. Her research interest lies in Computer Vision, Fashion Landmark Estimation and Human pose Estimation. She has already published the paper in ICASSP2020.

Hong Liu (M'08) received the Ph.D. degree in mechanical electronics and automation in 1996. He is currently a Full Professor in the School of Electronics Engineering and Computer Science, Peking University (PKU), Beijing, China. He has been selected as Chinese Innovation Leading Talent supported by "National High-level Talents Special Support Plan" since 2013. He is also the Director of Open Lab on Human Robot Interaction, PKU. He has published more than 150 papers. His research interests include computer vision and robotics, image processing, and pattern recognition. He received the Chinese National Aero-space Award, the Wu Wenjun Award on Artificial Intelligence, the Excellence Teaching Award, and the Candidates of Top Ten Outstanding Professors in PKU. He is the Vice President of Chinese Association for Artificial Intelligent (CAAI), and the Vice Chair of Intelligent Robotics Society of CAAI. He has served as keynote speakers, Co-Chairs, Session Chairs, or PC members of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC, and IIHMSP, and serves as reviewers for many international journals such as Pattern Recognition, the IEEE Transactions on Signal Processing, and the IEEE Transactions on Pattern Analysis and Machine Intelligence.

Wei Shi received the B.E. degree in electronic information engineering in 2016, and is working toward the Ph.D. degree in the School of EE&CS, Peking University (PKU), China. His research interests include person re-identification, person search and computer vision. He has already published articles in ICASSP2018, ICIP2018, ICIP 2019, ICASSP 2020, Chinese Journal of Electronics and The Journal of Engineering.



Xia Li received his bachelor's degree from Beijing University of Posts and Telecommunications (BUPT) in 2017. He is currently pursuing his master's degree at Peking University (PKU), China, advised by Prof. Hong Liu and Prof. Zhouchen Lin. His research interest lies in Computer Vision, Semantic Segmentation and Low-level Vision. He has published papers on academic conferences including CVPR' 20, AAAI'20, ICCV'19, MICCAI'19 and ECCV'18.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: