applied sciences

MDPI

Article

P D N: A Priori Dictionary Network for Fashion Parsing

Jue Hou [1,2], Yinwen Lu [1], Yang Yang [1,2] and Zheng Liu [2,3,*]

[1] School of Fashion Design & Engineering, Zhejiang Sci-Tech University, No. 928, 2nd Avenue, Qiantang District, Hangzhou 310018, China; hj1990@zstu.edu.cn (J.H.); 2021204010008@mails.zstu.edu.cn (Y.L.); yangy@zstu.edu.cn (Y.Y.)
[2] Key Laboratory of Silk Culture Heritage and Products Design Digital Technology, Ministry of Culture and Tourism, Hangzhou 310018, China
[3] School of International Education, Zhejiang Sci-Tech University, No. 928, 2nd Avenue, Qiantang District, Hangzhou 310018, China
* Correspondence: koala@zstu.edu.cn

Abstract: The task of fashion parsing aims to assign pixel-level labels to clothing targets; thereby, parsing models are required to have good contextual recognition ability. However, the shapes of clothing components are complex, and the types are difficult to distinguish. Recent solutions focus on improving datasets and supplying abundant priori information, but the utilization of features by more efficient methods is rarely explored. In this paper, we propose a multi-scale fashion parsing model called the Priori Dictionary Network (PDN), which includes a priori attention module and a multi-scale backbone. The priori attention module extracts high dimensional features from our designed clothing average template as a priori information dictionary (priori dictionary, PD), and the PD is utilized to activate the feature maps of a CNN from a multi-scale attention mechanism. The backbone is derived from classical models, and five side paths are designed to leverage the richer features of local and global contextual representations. To measure the performance of our method, we evaluated the model on four public datasets, the CFPD, UTFR-SBD3, ModaNet and LIP, and the experimental results show that our model stands out from other State of the Art in all four datasets. This method can assist with the labeling problem of clothing datasets.

Keywords: fashion parsing; attention mechanism; priori knowledge; convolutional neural network

Citation: Hou, J.; Lu, Y.; Yang, Y.; Liu, Z. P D N: A Priori Dictionary Network for Fashion Parsing. *Appl. Sci.* **2024**, *14*, 3509. https://doi.org/10.3390/app14083509

Academic Editor: Antonio Fernández-Caballero

Received: 25 March 2024
Revised: 13 April 2024
Accepted: 19 April 2024
Published: 22 April 2024

1. Introduction

As a fine-grained image semantic segmentation problem, fashion parsing aims to predict the label (e.g., T-shirt, bag, face, etc.) for each pixel in a fashion image. It has garnered unprecedented attention due to the fast growth of fashion-related applications, such as virtual try-ons and clothes recommendations [1,2]. In particular, models for high-level tasks typically use parsing results as input conditions, which benefit the precision of feature extraction [3,4]. However, labeling each pixel is a challenging task because some clothes with similar appearances have different categories in fashion parsing datasets, as shown in Figure 1. Therefore, the segmentation models cannot reliably forecast classifications due to the presence of confusing semantic information. In addition, the visual variations of targets, such as resolution, deformities, occlusions and background, impose great challenges for segmentation [5,6]. Early work on the fashion parsing problem usually adopted the modified image segmentation models, e.g., Grab Cut and the Markov Random Field (MRF). These models always greatly benefit from incorporating shape and texture features, but they can only process limited types of clothing [7,8].

Figure 1. Samples from fashion parsing dataset, in which different colors of ground truth represent different categories.

Recently, convolutional neural networks (CNNs) have made remarkable progress in image segmentation tasks, primarily due to their strong capabilities of feature capturing [9–12]. The prevalent CNN models for fashion parsing were based on some classical end-to-end networks, e.g., fully convolutional networks (FCNs) and U-Net [13,14], and they focused on improving the dataset quality to acquire more accurate results. However, dataset imbalance and low annotation quality are very common in human parsing problems in some widely used datasets, such as ModaNet [15] and LIP [16]. Therefore, these methods are difficult to adopt in practical applications.

Considering the limitation of the feature representation ability of classical CNN models, Vozarikova et al. [17] designed extra branches to learn richer features from images and replaced the backbone of U-Net with ResNet-34. Zhu et al. [18] proposed a progressive cognitive structure to segment human parts, in which the latter layers inherit information from the former layers to improve the recognition ability of small targets. To help the CNN models explicitly learn important information (e.g., edges, poses), some approaches attempted to extract priori knowledge from images and introduce them into models [19,20]. Ihsan et al. [21] injected superpixel features, which were extracted from the Simple Linear Iterative Clustering (SLIC) algorithm, into the decoder path of a modified FCN. However, their method needs post-processing steps to extract the templates of clothing, which means it is not an end-to-end model. Different from manually extracting features, Ruan et al. [22] designed a side branch for an edge detection task, which aims to improve the continuity of the segmentation results using the features of edges. Liang et al. [16] proposed to build two networks, where the networks were utilized to learn the mask templates and texture features, respectively. Park et al. [23] designed a tree structure to help the CNN model infer human parts in the hierarchical representation. For using human poses to provide object-level shape information, Gong et al. [24] used poses to directly supervise the parsing results and collected more than 50,000 images to build an evaluation dataset. Xia et al. [25] proposed a human parsing approach that uses human pose location as cues to provide pose-guided segment proposals for semantic parts. Zhang et al. [26] designed a five-stage model to extract hierarchical features from edges and poses and leverage them to generate parsing results. However, the extracted edges and poses always contain errors, which will be accumulated in the final results.

More recently, researchers found that highlighting relevant features using the self-attention mechanism is another method to improve fashion parsing models [27,28]. He et al. [29] incorporated three attention modules into a CNN, and they found that strengthening the connections between pixels using self-attention modules can increase clothing

parsing accuracy. Therefore, one natural question for using a CNN model to solve the problem of fashion parsing is, can we extract priori knowledge and introduce it into a CNN model using an attention mechanism?

To address these issues, we propose a novel framework aimed at leveraging priori attention with fewer errors to guide the generation of fashion parsing results. Considering that the explicitly extracted priori knowledge (edges and poses) always involves errors, we implicitly extracted latent features for each category using an encoding model. Then, the non-linear combinations of features were utilized to construct a series of category-based templates. We stored these latent space templates in the priori dictionary (PD). To employ the PD in guiding the backbone of a fashion parsing network, a modified attention module is proposed. The entire model is named the Priori Dictionary Network (PDN). The PDN takes advantage of feature templates and an attention mechanism to generate segmentation results, and its contributions are threefold: (1) we put forward a priori dictionary module to help the CNN model capture important features using templates, in which the module combines the advantage of priori information and an attention mechanism; (2) we proposed a joint loss function, L2 loss and cross-entropy loss to supervise the model training, which made the accuracy and IoU increase by 0.45% and 6.01%, respectively; and (3) we developed a multiple path architecture to fuse clothing features on a multi-scale, and the comparison results show that the proposed model outperforms State-of-the-Art methods in public fashion parsing datasets.

2. Methods

In this section, we will describe our method of how to acquire high-quality segmentation results from fashion parsing datasets.

2.1. Priori Dictionary Construction

Let us consider that $\mathcal{F}(x)$ is a CNN model, and it can be represented as follows:

$$\mathcal{F}(x) = \xi(\psi(x)) \tag{1}$$

where ψ denotes the convolutional layers and pooling layers of the CNN model, and ξ denotes the fully connected layers. We assume that $\mathcal{G}(x)$ is a segmentation of the CNN model, and the backbone of the model $\mathcal{G}(x)$ has the same structure as ψ. Hence, the model $\mathcal{G}(x)$ can easily be represented as:

$$\mathcal{G}(x) = \rho(\psi'(x)) \tag{2}$$

where the ρ indicates the rest of the part of $\mathcal{G}(x)$. If there is a dataset Ω that involves classification labels and pixel-wise annotations (e.g., the CFPD and ModaNet), we train $\mathcal{F}(x)$ and $\mathcal{G}(x)$ on the dataset Ω, and two groups of features can be extracted by ψ and ψ', respectively. We let Δ denote the differences between the two groups of features:

$$\Delta = \psi'(x) - \psi(x) \tag{3}$$

Then, the partial derivative of $\psi'(x)$ can be rewritten as follows:

$$\frac{\partial \psi'}{\partial x} = \frac{\partial \Delta}{\partial x} + \frac{\partial \psi}{\partial x} \tag{4}$$

Equation (4) represents that the CNN model can more easily approximate the desired function (the segmentation problem) using priori knowledge. Indeed, humans always prefer to utilize priori knowledge to tackle complicated tasks. For example, the semantic information of clothes (e.g., texture and clothing shapes) is important evidence for the manual labeling pixels of fashion parsing maps, and it is highly related to the clothing category. Therefore, the clothing category can also help the CNN to understand objects properly and be regarded as a priori condition for parsing map prediction.

2.1.1. Priori Feature Learning

Latent variables are useful for capturing complex or conceptual properties of inputs [30]. To guide the feature extraction of our fashion parsing model, we utilized a high-accuracy classification model to build a series of feature templates, and the templates were adopted to enhance the crucial features in latent space. Assume that the dataset Ω is a fashion dataset, and it involves N classes of items, such as skirts, glasses and sweaters. Let Ω_n denote the nth subsets of Ω, and it is composed of the same kind of items. Thus, we can train a CNN model $\mathcal{G}$ on the dataset Ω to predict the category of item in the image. After the model convergence, it is easy to extract latent representations L_{ni} according to the categories using correct predictions, where L_{ni} indicates the features of the i th image in Ω_n.

Most of the CNN models employ an encoder structure to extract features, and their backbones always group several convolutional layers and one pooling layer as a stage, such as VGG-16, ResNet-50 and ResNet-101. For example, the VGG-16 model groups 13 convolutional layers to 5 stages, and the latent representations L_{ni} of VGG-16 are composed of 13 feature maps $L_{ni} \in \left\{ l_{ni}^{(1)}, l_{ni}^{(2)}, \ldots l_{ni}^{(13)} \right\}$. Then, we preserved the last output features of each stage, which means L_{ni} reserved the feature maps of the 2nd, 4th, 7th, 10th and 13th convolutional layers.

To obtain the general representations of items, we computed the feature templates for each class as follows:

$$T_n = \frac{\sum_{i=1}^{k} (L_{ni}; \mathcal{G}(x_{n,i}))}{k} \tag{5}$$

where T_n denotes the feature templates of Ω_n, and k represents the object number of categories. The function $\frac{\sum_{i=1}^{k} (L_{ni}; \mathcal{G}(x_{n,i}))}{k}$ of Equation (5) was employed to average the feature maps of different objects. As shown in Figure 2, we fused the channels of acquired templates and visualized them. The visualization results indicate that the feature templates involve high-response regions corresponding to their categories, and the pixel-wise labeling model can benefit from the templates (see more details in Section 2.2.1). To store the templates $T \in \{T_1, T_2, \ldots, T_n\}$, we built a priori dictionary (PD), as illustrated in Algorithm 1.

As shown in Figure 3, our priori dictionary essentially is a 4D matrix, and it consists of N groups of templates. It is worth noting that for different model structures, the latent representations used to construct the PD are sourced from different convolutional layers. In this research, the reserved latent representations of ResNet-101 were generated by the 1st, 10th, 22nd, 91st and 100th convolutional layers.

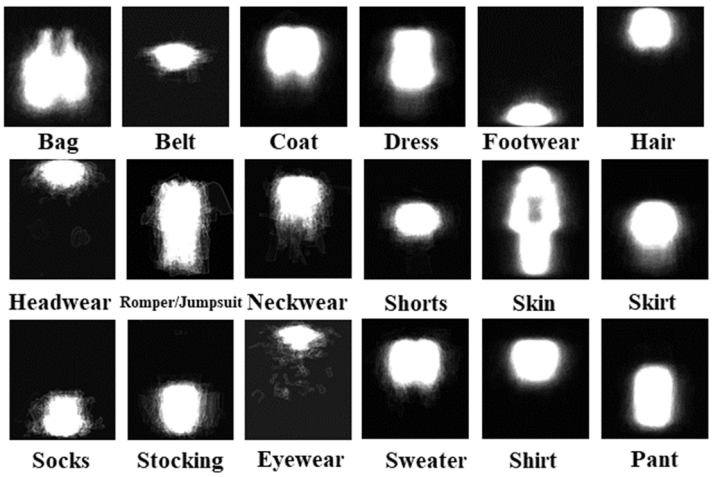

Figure 2. The visualization results of latent feature templates.

Algorithm 1 Generation of priori dictionary for VGG-16

$\{\Omega_1, \Omega_2, \dots, \Omega_N\} \in \{\{x_{11}, x_{12}, \dots, x_{1I}\}, \dots, \{x_{N1}, x_{N2}, \dots, x_{NI}\}\}$ ~ $d$ataset for $\mathcal{G}$ training
When $\mathcal{G}$ convergence do
  for $n = 1, 2, \dots, N$
    select k images from Ω_n
    for $i = 1,2, \dots, k$
      $\left\{l_{ni}^{(1)}, l_{ni}^{(2)}, \dots, l_{ni}^{(13)}\right\} = \mathcal{G}(x_{ni})$ ~ generate latent representations for x_{ni}
      $\left\{l_{ni}^{(1)}, l_{ni}^{(2)}, \dots, l_{ni}^{(13)}\right\} \rightarrow \left\{l_{ni}^{(2)}, l_{ni}^{(4)}, l_{ni}^{(7)}, l_{ni}^{(10)}, l_{ni}^{(13)}\right\}$ ~ reserve latent representations
    end
    $T_n^{(1)} = \frac{\Sigma_k l_{ni}^{(2)}}{k}$, $T_n^{(2)} = \frac{\Sigma_k l_{ni}^{(4)}}{k}$, $T_n^{(3)} = \frac{\Sigma_k l_{ni}^{(7)}}{k}$, $T_n^{(4)} = \frac{\Sigma_k l_{ni}^{(10)}}{k}$, $T_n^{(5)} = \frac{\Sigma_k l_{ni}^{(13)}}{k}$
      ~ compute templates of feature maps
  end
  $D \in \left\{\left\{T_1^{(1)}, T_1^{(2)}, T_1^{(3)}, T_1^{(4)}, T_1^{(5)}\right\}, \dots, \left\{T_n^{(1)}, T_n^{(2)}, T_n^{(3)}, T_n^{(4)}, T_n^{(5)}\right\}\right\}$
    ~ build the priori dictionary D
Output $D \in \{D^{(1)}(T_1^{(1)}, T_2^{(1)}, \dots, T_n^{(1)}), \dots, D^{(5)}(T_1^{(5)}, T_2^{(5)}, \dots, T_n^{(5)})\}$
    ~ reshape the priori dictionary D
end

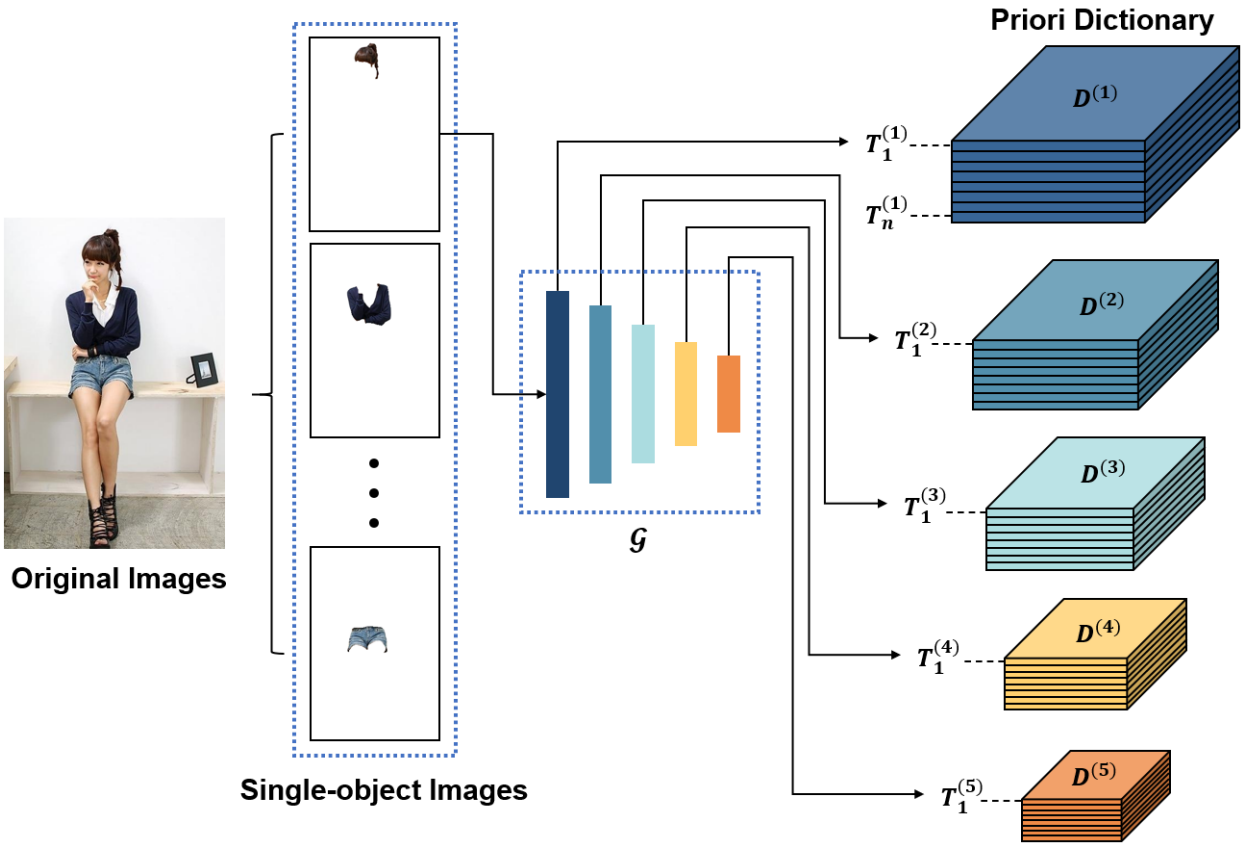

Figure 3. An illustration of PD construction. The latent representations are extracted from the CNN, and different colors indicate that the representations are extracted from different kinds of objects.

2.1.2. Dataset Reconstruction

As shown in the illustration in Section 2.1.1, the CNN model is trained on classification datasets to build the PD, and the features are stored based on the corresponding categories. However, the images of fashion parsing datasets are collected from the real world, and

there are several kinds of targets in one image. In addition, the images are annotated with pixel-wise labeling to evaluate clothing parsing, e.g., the CFPD [31] and Fashionista [32]. Therefore, the original datasets cannot be utilized in the classification task. Here, we formulate the dataset of fashion parsing as:

$$\Omega(x,y) = \Omega_1(x,y) \bigcup \Omega_2(x,y) \bigcup \ldots \bigcup \Omega_n(x,y) \tag{6}$$

where $\Omega(x,y)$ represents all the pixels in images, and $\Omega_n(x,y)$ denotes the set of pixels of the nth image. Aiming to build a dataset for the classification task, we built a new single-object dataset according to the class annotations. Compared with real-world images, single-object images remove the backgrounds and reserve the clothing pixels. Notably, one real-world image was deconstructed into several single-object images so that the single-object image only included one kind of clothing, as shown in Figure 2. The reconstructed dataset can be formulated as follows:

$$\Omega(c_1, c_2, \ldots, c_n) = \Omega_{c1} \bigcup \Omega_{c2} \bigcup \ldots \bigcup \Omega_{cn} \tag{7}$$

where $\Omega(c_1, c_2, \ldots, c_n)$ is the reconstructed dataset, and Ω_{cn} indicates the set of the nth class of items. Therefore, object-level categories can be utilized to build a PD.

2.2. Multi-Scale Feature Fusion Model

2.2.1. Priori Attention Module

Attention mechanisms improve the model performance on semantic analysis and image processing by capturing long-range feature interactions and enhancing the representation of features for a CNN [33–35]. To utilize the classification and shape information of a PD, we designed an attention-based module to introduce the templates into a fashion parsing model, namely, the priori attention module.

As shown in Figure 4, we applied convolutional kernels to fuse the priori dictionary and acquire a series of weighted templates. Specifically, the jth stage of the priori dictionary $D^{(j)}$ comprises c (number of categories) groups of feature templates with d channels, where d is equal to the channel number of features from the backbone. We used c convolutional kernels to fuse $D^{(j)}$, and the dimension of weighted templates is equal to $c \times h \times w$, where h and w represent the height and width of the feature maps. Since the convolutional kernels were employed to fuse $D^{(j)}$ independently, each channel of the weighted templates still corresponded to one category. Then, the 3D weighted templates were reshaped to a 2D matrix as a query vector.

For the feature maps of the backbone, two convolutional kernels were adopted to fuse them, and the dimensions of the fused matrixes are equal to $c \times h \times w$. Then, we reshaped two matrixes to a pair of 2D matrixes, which are named as the value vector and the query vector. Different from the self-attention mechanism, the key vectors and query vectors were generated by different sources. To utilize the priori attention to guide the feature extraction of the CNN, we performed a softmax function to activate the multiplication results of the query vector and key vector as follows:

$$Attention(Q,K) = softmax\left(QK^T\right) \tag{8}$$

where Q and K are the query and key vectors. It is worth noting that we obtained a matrix composed of attention values using multiplication, and the features that could be matched with the templates could receive higher response values. Then, we used pixel-wise multiplication and addition to enhance the important features of the value vectors and reshaped them to a c-channel feature map. At the end of the priori attention module, a convolutional layer was employed to restore the dimension of the feature map.

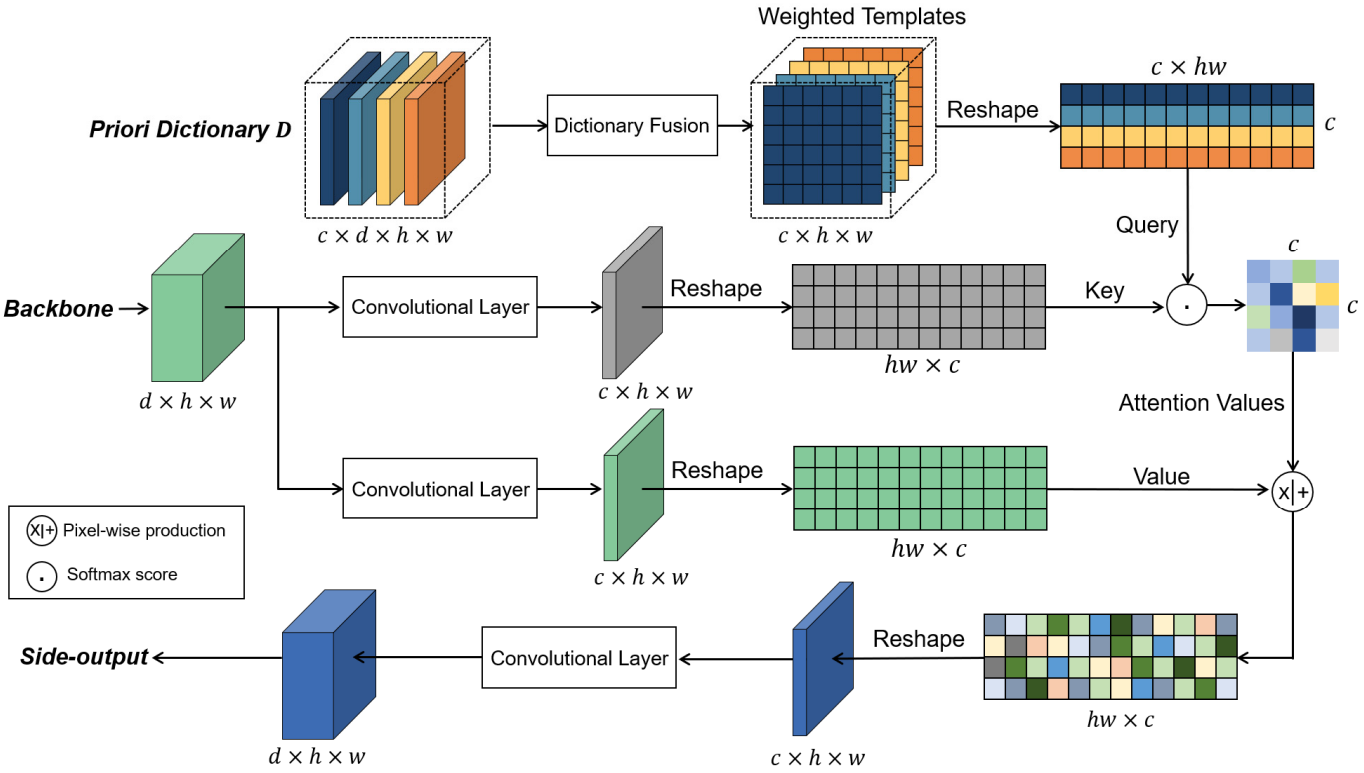

Figure 4. The illustration of attention module. Two reshaped vectors (query vector and key vector) multiply with each other and create the attention matrix, which is used to enhance the useful features. At the end of attention module, a convolutional kernel is applied to restore the feature maps whose dimensions are equal to the dimensions of input feature maps.

2.2.2. Network and Side Paths for Multi-Scale Feature Fusion

The receptive field limits the range of features that convolutional kernels can typically extract, posing challenges for leveraging both global features and local detail features. Therefore, through the design of side paths, we utilized the semantic information of global features to determine the position and shape of the target and detailed features to ensure the continuity and integrity of the same target.

As shown in Figure 5, we added a priori attention module behind each convolutional block. Then, we duplicated the outputs of the priori attention modules and fed them into the side paths. In each side path, we used a $1 \times 1 \times 64$ convolutional kernel to obtain a feature map with 64 channels and deconvolutional layers to restore their heights and widths. At the end of the network, we concatenated all the feature maps from the side paths and designed a fuse block comprising three convolutional layers. Notably, the last convolutional layer outputs the predicted results with the c-channel, where each channel represents the probability map for one category. Since the Conv Blocks of our model were inherited structures from widely used models (VGG-16 and ResNet-101), we easily inferred the parameters of the priori attention modules and side paths based on the used dataset.

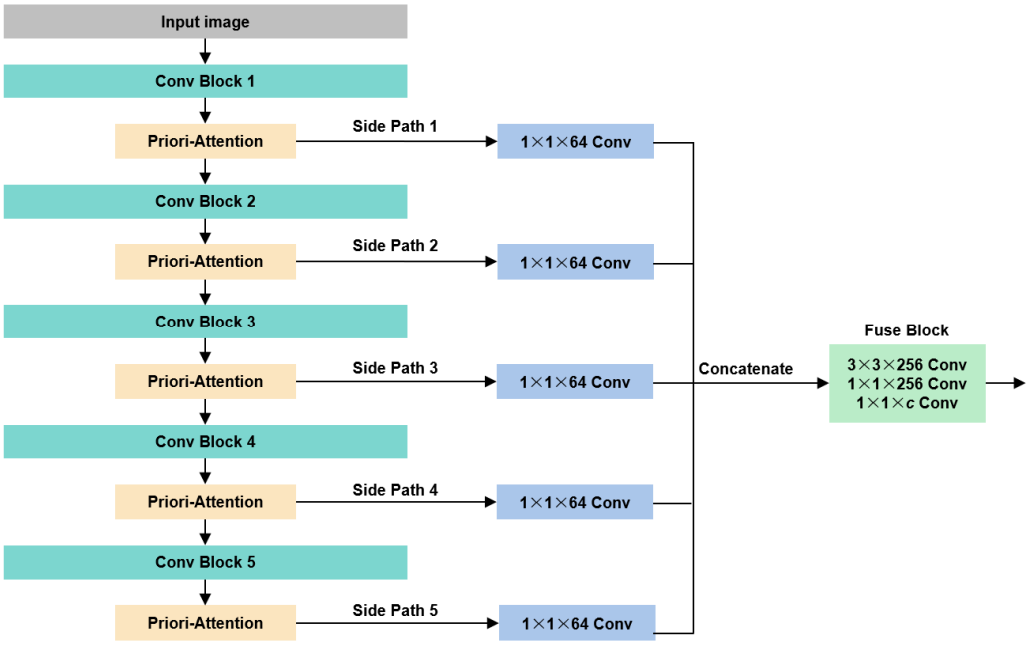

Figure 5. The illustration of PDN, in which the backbone uses VGG16 as an example.

2.2.3. Joint Loss Function

A fashion parsing model is designed to acquire a segmentation map that involves pixel-level labels. Therefore, the fashion parsing problem can be regarded as two subproblems: pixel classification forecasting and semantic segmentation. To solve both of the subproblems, we propose a cross entropy and L2 joint loss function as follows:

$$Loss = L_{cross} + L_2 = \frac{1}{c}\sum_{i,j}\sum_{k=1}^{c} y_{i,j,k} log\left(P_{i,j,k}\right) + \frac{\sum_{1}^{W\times H}\left|P_{i,j,k} - y_{i,j,k}\right|}{W \times H} \quad (9)$$

where H, W and c are the height of images, width of images and category count of datasets, respectively. The outputs of our model were reshaped into a matrix whose dimensions are equal to $h \times w \times c$, where each channel is a binary map and only includes one kind of target prediction. The L_2 term was designed for optimizing all the prediction results in all channels, and the cross-entropy loss was proposed to optimize the shapes of targets in each channel.

3. Experiments

We conducted experiments on four publicly available datasets: the CFPD [31], UTFPR-SBD3 [5], ModaNet [15] and LIP [16]. We implemented the proposed model using the public deep learning library Pytorch on a single RTX-3090 GPU. We extracted single-target images from the four datasets mentioned above and constructed priori dictionaries for each of them. The details and results are illustrated in the following sections.

3.1. CFPD and UTFPR-SBD3 Datasets

The Colorful Fashion Parsing dataset (CFPD), which contains 23 types of categories, is widely used for image parsing evaluation. In this dataset, there are 2682 images whose heights and weights are equal to 600 and 400, respectively. However, the annotations of the CFPD are heavily influenced by noise, and their incorrect labels affect the accuracy of CNN models. Thus, Inacio et al. collected images from the CFPD dataset and Refined Fashionista dataset and manually annotated the pixel-level categories to obtain high-quality ground truth, namely, the Soft Biometrics dataset (UTFPR-SBD3). The UTFPR-SBD3 dataset includes 17 types of categories and 4500 images, whose resolutions are also equal to 600×400.

3.1.1. Implementation Details

PD construction of the CFPD dataset. To test our model on the CFPD dataset, we adopted VGG-16 as the backbone of the PDN model. After the reconstruction of the single-object dataset of the CFPD, we trained the VGG-16 model to build the PD. VGG-16 was initialized with the parameters pretrained on ImageNet, and the Adam Optimizer was adopted. We set the learning rate to 10^{-5}, and it was divided by 10 after every five epochs. The training was terminated after 20 epochs on the CFPD when the loss and accuracy of the validation set could not change. We tested VGG-16 on the testing set of the reconstructed dataset, and the accuracy of classification reached 98.2%. Then, we randomly selected 30 single-object images for each category from the training set and computed the feature templates of the PD.

PD construction of the UTFPR-SBD3 dataset. The VGG-16 network adopted similar hyper-parameters as the training on the CFPD dataset. We found that the model convergences after 25 epochs of training, and the accuracy reaches 99.1%. Similar to the CFPD dataset, each template of the PD was also obtained using 30 single-object images.

Full model training. According to Reference [5], UTFR-SBD3 and the CFPD involve part of the same image for evaluation. Therefore, we adopted a similar network structure and training parameter for them. Specifically, we initialized the backbone (VGG-16) with the parameters pretrained on ImageNet, and the convolutional layers of the side paths and fuse block were initialized using the Xavier method [36]. The Adam Optimizer was also adopted, and the models converged at 20 epochs in both datasets. In addition, the batch size was set to 16, and the learning rate was fixed at 10^{-5}.

3.1.2. Results

For a fair comparison, we used the same measures (Acc and mIoU) reported by [5] to evaluate the parsing quality of the CFPD dataset. The experimental results are reported in Table 1. It can be seen that the accuracy and IoU of our model achieve 99.26% and 78.09%, respectively. Among a series of SOTA methods, EPYNET achieves the best performance, and our method further improves the accuracy by 5.8% and mIoU by 31.2% compared to EPYNET.

Table 1. Results of experiments on CFPD dataset.

Reference	Method	Accuracy	mIoU
[31]	Color-Fashion	-	42.10
[21]	Superpixel Parsing	88.69	-
[13]	Weak Structural	93.06	53.51
[37]	Locality Aware	93.52	53.00
[38]	Feature Pyramid	93.82	54.39
[5]	EPYNET	93.83	59.52
[32]	PaperDoll	89.89	-
[39]	ATR	98.67	-
	PDN (ours)	99.26	73.99

Compared with the CFPD dataset, UTFR-SBD3 is a more balanced dataset, and it has higher-quality annotations. We evaluated our model on the test set of UTFR-SBD3, and the PDN achieves 99.39% accuracy and 81.00% mIoU, respectively, which are 18.1% and 15.4% higher than the EPYNET. The results show that the higher quality of annotations can help our model show a better performance. Figure 6 shows the confusion matrixes of the PDN and EPYNET on UTFR-SBD3, and we found that our model not only achieves a better overall performance but also more accurate predictions for all categories.

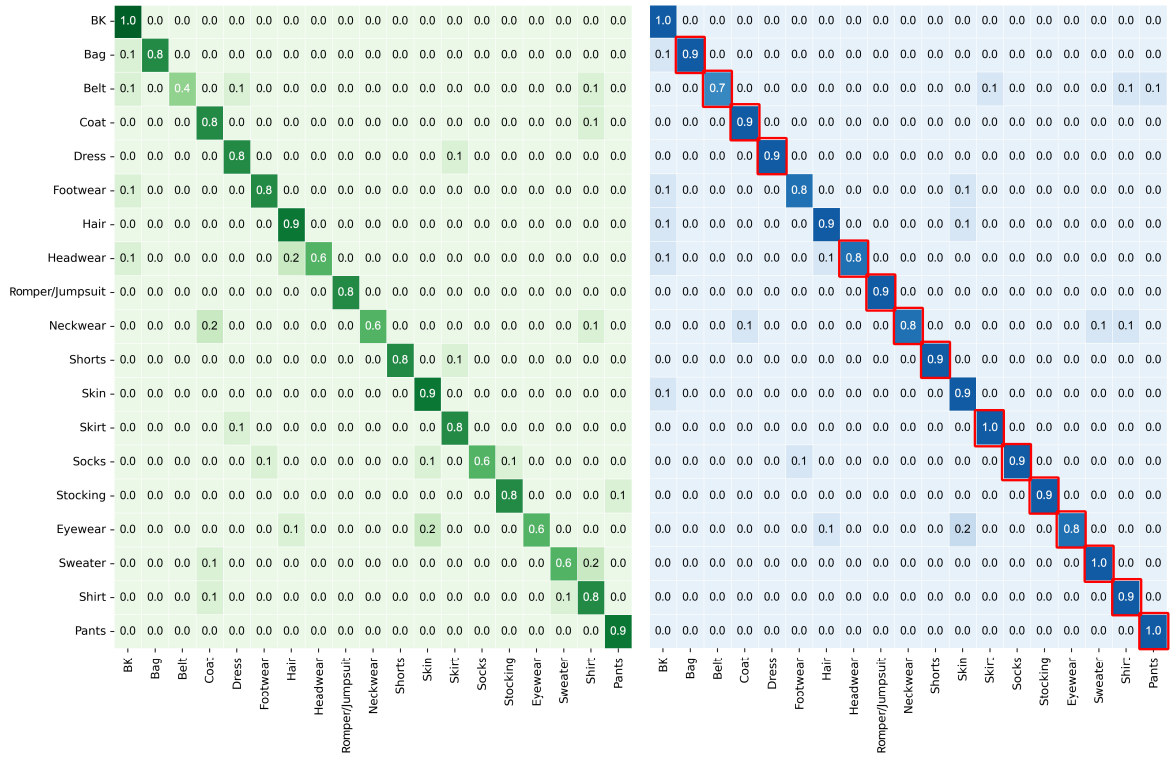

Figure 6. The confusion matrix of PDN results on the UTFR-SBD3 dataset, in which the left and right are the result of EPYNET and PDN, respectively.

In Figure 7, a comparative visualization of the PDN algorithm on the UTFR-SBD3 dataset is presented. The first column displays input images, the second column shows the corresponding ground truth, the third column exhibits the segmentation results obtained using the EPYNET method, and the fourth column showcases the segmentation results obtained using the PDN method.

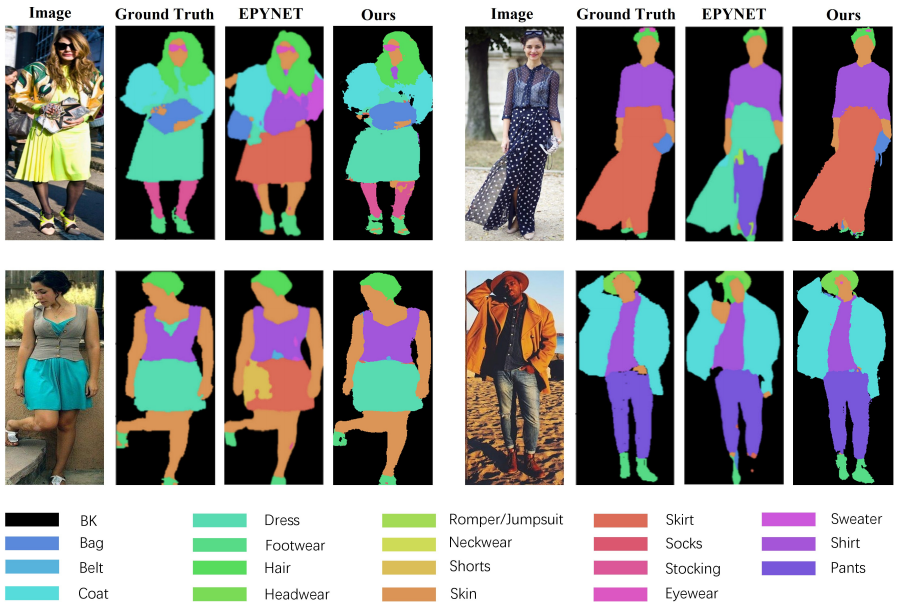

Figure 7. Samples of PDN results. The first column includes original images from UTFR-SBD3 dataset, the second column includes ground truth, the third column includes the results of EPYNET, and the fourth column includes the results of PDN.

It can be observed that the PDN is more effective in recognizing correct categories and shapes. In the first set of images, the PDN successfully identifies the category of the skirt, whereas EPYNET misclassifies it. In the second and third sets of images, EPYNET segments the lower garment into two parts, namely, dress and pants, whereas the PDN closely approximates the ground truth. In the fourth set of images, EPYNET fails to completely segment the shoes, whereas the PDN achieves a complete segmentation result. These significant outcomes stem from the multi-scale architecture and attention modules designed within the PDN.

3.2. ModaNet Dataset

ModaNet is the largest dataset of fashion parsing [14]. Compared with the UTFPR-SBD3 dataset, the human parsing dataset (ModaNet) involves fewer kinds of clothing labels. Specifically, ModaNet involves 55,176 street images, and the images are annotated by a polygonal mask to 13 categories. We divided the training set (52,254 images) into a training set, validation set and testing set, where the subsets involve 45,000, 5000 and 2254 images, respectively.

3.2.1. Implementation Details

PD construction of ModaNet. We trained VGG-16 on ModaNet, and the results show that the accuracy and IoU only achieve 81.78%. As ModaNet involves more low-quality images, the rough annotations mean that the VGG-16 model cannot extract features. Therefore, we replaced the VGG-16 model with the ResNet-101 model for extracting priori features. ResNet-101 was trained on the reconstructed dataset, where the model was initialized with ImageNet-pretrained ResNet-101. After 20 epochs of training, the model achieved 93% accuracy. Similar to the CFPD and UTFR-SBD3, we also randomly selected 30 images for each category from the training set and generated the PD.

Full model training. Considering the poor performance of the VGG-16 model on the ModaNet dataset, we used ResNet-101 as the backbone of the full model. We initialized the convolutional blocks with ImageNet-pretrained ResNet-101, and other convolutional layers were initialized using the Xavier method. During training, Adam was chosen as the optimizer for our model. We set the learning rate and batch size to 10^{-4} and 16, respectively, and the training was terminated after 40 epochs.

3.2.2. Results

As shown in Table 2, we compared our model with other SOTA methods, and the PDN with the ResNet-101 backbone yielded a result of 77.91% in mIoU, which is 8.36% higher than TANet. To validate the effectiveness of the PD, we also trained the PDN with a VGG-16 backbone. Although this weaker model only achieved 44.80% in mIoU, it still achieved a superb performance among a series of VGG-16-based models. This result unveils that our proposed design brings great benefit to the CNN models in fashion parsing tasks.

Table 2. Results of experiments on ModaNet dataset.

Reference	Method	Backbone	mIoU
[31]	FCN-32s	VGG	35.36
[40]	FCN-16s	VGG	36.93
[13]	FCN-8s	VGG	38.00
[41]	DANet	ResNet-101	68.16
[42]	Deeplabv3	ResNet-101	68.57
[43]	PSPNet	ResNet-101	68.80
[44]	CCNet	ResNet-101	69.22
[29]	TANet	ResNet-101	69.55
	PDN (VGG-16)	VGG	44.80
	PDN (ResNet-101)	ResNet-101	77.91

We compared the per-class IoU of our model with TANet, and the results are listed in Table 3. The comparison results show that the PDN can generate higher quality segmentation maps for most of the classes except belts and sunglasses, since belts and sunglasses only occupy 6.0% and 3.8% of all the targets, respectively, and they tend to be smaller compared to other types of targets. It is difficult to capture sufficient features for these small objects, leading to difficulties for the PD in storing templates of the latent representations. However, when the backbone can capture sufficient features, our method still significantly improves fashion parsing accuracy.

Table 3. Per-class IOU on ModaNet.

Class	TANet (IoU%)	PDN (ResNet-101) (IoU%)
Background	98.38	98.33
Bag	77.49	79.24
Belt	55.28	40.55
Boots	50.97	81.39
Footwear	56.35	69.34
Coat/Jacket	73.0.7	85.81
Dress	70.23	90.79
Sunglass	61.56	44.96
Pants	81.41	91.69
Top	72.05	85.67
Shorts	73.60	83.72
Skirt	71.50	92.65
Headwear	73.35	74.22
Scarf/Tie	58.42	69.88

In Figure 8, we present the testing results on the ModaNet dataset. The first column shows input images, the second column shows ground truth, and the third column shows the segmentation results obtained using the PDN method. It is evident that the PDN achieves segmentation results closest to the ground truth, both in terms of category classification accuracy and shape segmentation similarity. Particularly noteworthy is its performance on components such as scarves or bags, where the PDN exhibits notable accuracy.

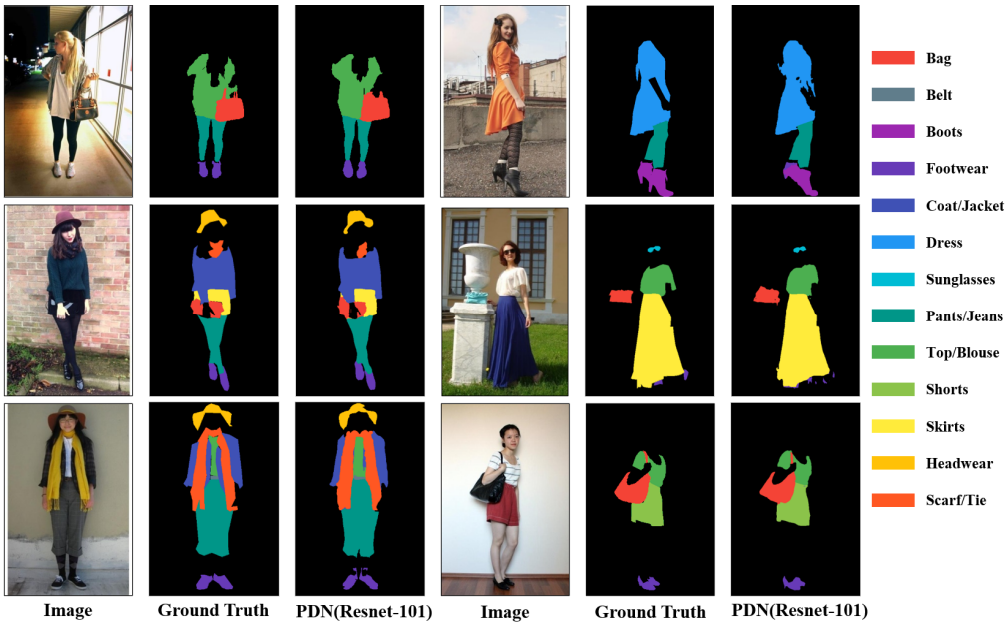

Figure 8. The testing samples of PDN on ModaNet dataset.

3.3. LIP Dataset

The LIP dataset is the most widely used human parsing dataset, with 19 types of categories. This dataset contains 50,462 images, including 13,672 upper-body images, 19,081 full-body images, 3386 head-missed images and 2778 back-view images. We followed the previous approaches [16] and divided the dataset into 30,462 training images, 10,000 validation images and 10,000 test images. Aiming to explore the performance of the PDN in low-data scenarios, we randomly selected 3046 images to construct a separate training set with limited data. To avoid misunderstanding the notations, let "training set" and "training set (3K)" denote the original training set and the limited training set.

3.3.1. Implementation Details

PD construction of LIP. Considering the poor performances of the VGG-16 model in feature extraction, we adopted ResNet-101 to build the priori dictionary on the LIP dataset. Specifically, the ResNet-101 model was trained on the single-object dataset with the Adam optimizer. The learning rate and batch size were set to 10^{-5} and 16, respectively. After 35 epochs of training, the accuracy of ResNet-101 achieved 89.7%. Then, we computed the feature templates to build the PD, in which each template was also obtained by randomly selecting 30 images. Moreover, for a fair comparison, we adopted the same parameters to train ResNet-101 on the reconstructed single-object dataset of the training set (3K) and generate a new priori dictionary.

Full model training. We adopted the same network architecture and hyper-parameters settings as those used to train the PDN on the ModaNet dataset. For the complete training set, the training was terminated after 35 epochs; for the training set (3K), the training was terminated after 22 epochs. In the following section, the PDN trained on the training set (3K) is noted as PDN-3K.

3.3.2. Results

For the LIP dataset, we used the LIP validation set as the test set because the official dataset does not provide annotation results for the test set. Similar to [45], we used the same evaluation matrices to measure the performance of our model, including pixel accuracy (PA), mean class pixel accuracy (MPA) and mean intersection over union (mIoU). As shown in Table 4, our model achieves the best PA, MPA and mIoU, which are equal to 76.02%, 54.46% and 45.12%, respectively. Due to the limited generalization caused by the scarcity of training data, the accuracy of PDN-3K noticeably decreased with low-data scenario training. However, it can be found that the evaluation results are still close to those of DeepLabv3+. Overall, our model still achieves the average performance level of segmentation algorithms, indicating that our proposed model exhibits good adaptability, even under extreme conditions. Additionally, the IoU metrics for each category are listed in Table 5. Our PDN model shows a more obvious improvement in upper clothes and skirts, and this is attributed to our proposed clothing priori dictionary module.

Table 4. Results of experiments on LIP dataset.

Reference	PA	MPA	mIoU
PSPNet [43]	74.95	54.20	44.22
DeepLabv3+ [42]	73.70	51.19	36.76
DANet [41]	72.06	54.11	36.51
DTPR [6]	74.10	45.92	44.29
PDN-3K (ours)	68.41	48.08	35.24
PDN (ours)	76.02	54.46	45.12

Table 5. Per-class IOU on the LIP dataset.

Class	PSPNet	DeepLabv3+	DANet	PDN
Hat	55.08	52.02	57.08	57.35
Hair	66.41	62.75	67.33	68.52
Glove	34.14	33.60	36.85	37.56
Sunglasses	31.49	26.06	31.17	31.69
Upper Clothes	65.59	64.32	65.77	67.85
Dress	20.77	20.37	18.66	19.64
Coat	51.41	49.34	50.92	52.17
Socks	40.43	37.68	41.23	39.39
Pants	71.05	72.70	71.27	71.19
Jumpsuits	26.96	27.94	28.14	29.56
Scarf	12.88	13.89	15.83	16.23
Skirt	15.59	14.64	16.09	17.54
Face	78.03	74.10	79.07	78.62
Left arm	57.84	59.73	59.15	60.09
Right arm	61.43	61.97	62.47	62.75
Left leg	47.93	48.36	47.90	48.03
Right leg	49.76	49.26	49.89	49.57
Left shoe	29.27	28.03	29.67	30.85
Right shoe	28.28	29.59	29.23	29.08
Avg	42.23	41.32	42.86	43.41

In Figure 9, we present a subset of the testing results on the LIP dataset, which are arranged from left to right as follows: the original images from the dataset, the results of PDN-3K, the results of the PDN and the ground truth. It can be observed that although there are some prediction errors in the PDN-3k model for categories such as Skirt and Shoe, the overall segmentation shape remains close to the ground truth, indicating that our model still exhibits good stability even in low-data scenarios. Moreover, the training results on the entire dataset closely approximate the ground truth across all categories, demonstrating that the proposed priori dictionary module can assist the model in capturing more accurate clothing details.

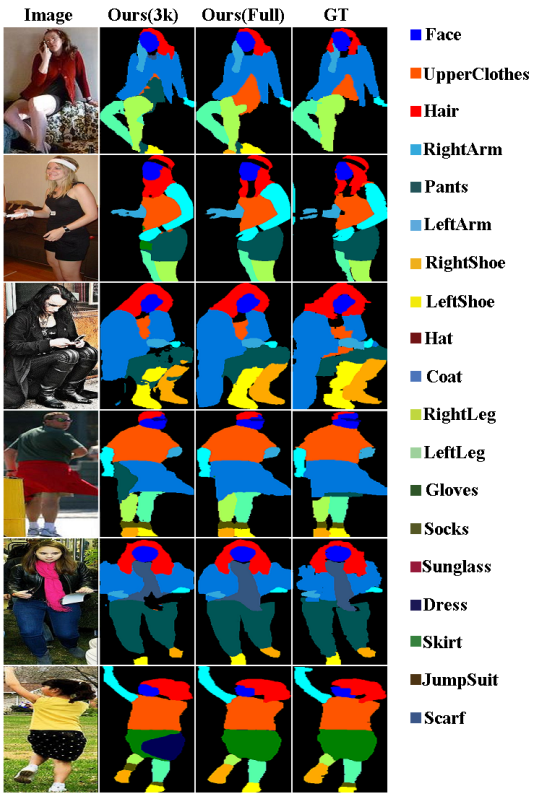

Figure 9. Testing samples of PDN on LIP dataset.

3.4. Ablation Study

We conducted ablation studies on the CFPD dataset to investigate the impact of different modules on the PDN. "Priori-attention" refers to the multi-scale attention mechanism, while "joint loss function" denotes the combined loss function used in this study's training. The results of the ablation experiments are presented in Table 6.

Table 6. Ablation experiments of each module. √ indicates that the module was used.

Priori Attention	Joint Loss Function	Accuracy	mIoU
		48.32	36.27
√		98.81	67.98
	√	59.57	25.39
√	√	**99.26**	**73.99**

Compared with the PDN, the model without a joint loss function acquires lower accuracy because the joint function focuses on both the accuracy of label assignment and the continuity of regions. The results show that the priori attention module can clearly improve performance. The model without the attention module and joint loss function cannot extract the segmentation maps of abundant targets, and the model only achieves 59.57% accuracy and 25.39% mIoU.

4. Conclusions

In this paper, we present a priori attention module-based CNN method for fashion parsing. First, we extracted features from training images and built a priori dictionary to store the features. Then, we attached the priori dictionary to a multi-scale backbone derived from VGG-16 and ResNet. Furthermore, a joint loss function that is composed of L2 loss and cross-entropy loss was proposed to guide the model training. We conducted our model on four public datasets, and the experiment results show that the CNN model can effectively annotate different types of targets by introducing our latent features, with mIOU values of 73.99, 79.91, 69.88 and 45.12 on the four datasets, respectively. Through a fair comparison with other State-of-the-Art models, we demonstrate that the PDN can achieve competitive results for fashion parsing. Additionally, the ablation study shows that the priori attention module and joint loss function improve the performance of the model. Although the PDN model can obtain precise parsing results on high-quality annotated datasets, the feature templates cannot lead to a significant improvement in the performance of predicting pixel labels for small targets. In future work, we will explore the use of priori knowledge introduction to improve parsing accuracy in more complex backbone models.

Author Contributions: Conceptualization, J.H. and Y.L.; methodology, J.H.; software, J.H. and Y.L.; validation, Y.Y.; formal analysis, Y.Y.; investigation, J.H. and Y.L.; resources, Z.L.; data curation, Y.L.; writing—original draft preparation, J.H.; writing—review and editing, J.H. and Y.L.; visualization, Y.Y.; supervision, Z.L.; project administration, Z.L.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Zhejiang Sci-Tech University of Technology Research Initiation Fund, grant number 21072325-Y.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Honda, S. Viton-gan: Virtual try-on image generator trained with adversarial loss. *arXiv* **2019**, arXiv:1911.07926.
2. Han, X.; Wu, Z.; Wu, Z.; Yu, R.; Davis, L.S. Viton: An image-based virtual try-on network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7543–7552.
3. Lu, Y.; Gu, B.; Ouyang, W.; Liu, Z.; Zou, F.; Hou, J. LSG-GAN: Latent space guided generative adversarial network for person pose transfer. *Knowl.-Based Syst.* **2023**, *278*, 110852. [CrossRef]
4. Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.-Y.; Lian, Z. Controllable person image synthesis with attribute-decomposed gan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5084–5093.
5. Inacio, A.D.S.; Lopes, H.S. EPYNET: Efficient pyramidal network for clothing segmentation. *IEEE Access* **2020**, *8*, 187882–187892. [CrossRef]
6. Li, Y.; Zuo, H.; Han, P. A Universal Decoupled Training Framework for Human Parsing. Sensors. *Sensors* **2023**, *22*, 5964. [CrossRef] [PubMed]
7. Boykov, Y.Y.; Jolly, M.-P. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In Proceedings of the Proceedings Eighth IEEE International Conference on Computer Vision, ICCV 2001, Vancouver, BC, Canada, 7–14 July 2001; pp. 105–112.
8. Gallagher, A.; Chen, T. Clothing cosegmentation for recognizing people. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 24–26 June 2008; pp. 1–8.
9. Bo, Y.; Fowlkes, C.C. Shape-based pedestrian parsing. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 2265–2272.
10. Chen, H.; Gallagher, A.; Girod, B. Describing clothing by semantic attributes. In Proceedings of the Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 609–623.
11. D'Angelo, A.; Dugelay, J.-L. Color based soft biometry for hooligans detection. In Proceedings of the 2010 IEEE International Symposium on Circuits and Systems, Paris, France, 30 May–2 June 2010; pp. 1691–1694.
12. Perlin, H.A.; Lopes, H.S. Extracting human attributes using a convolutional neural network approach. *Pattern Recognit. Lett.* **2015**, *68*, 250–259. [CrossRef]
13. Chen, Z.; Liu, S.; Zhai, Y.; Lin, J.; Cao, X.; Yang, L. Human parsing by weak structural label. *Multimed. Tools Appl.* **2018**, *77*, 19795–19809. [CrossRef]
14. Hrkac, T.; Brkic, K.; Kalafatic, Z. Multi-class U-Net for segmentation of non-biometric identifiers. In Proceedings of the 19th Irish Machine Vision and Image Processing Conference, Maynooth, Ireland, 30 August–1 September 2017; pp. 131–138.
15. Zheng, S.; Yang, F.; Kiapour, M.H.; Piramuthu, R. Modanet: A large-scale street fashion dataset with polygon annotations. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1670–1678.
16. Liang, X.; Gong, K.; Shen, X.; Lin, L. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 871–885.
17. Vozáriková, G.; Stana, R.; Semanisin, G. Clothing Parsing using Extended U-Net. In Proceedings of the VISIGRAPP (5: VISAPP), Online Streaming, 8–10 February 2021; pp. 15–24.
18. Zhu, B.; Chen, Y.; Tang, M.; Wang, J. Progressive cognitive human parsing. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
19. Liu, K.; Choi, O.; Wang, J.; Hwang, W. Cdgnet: Class distribution guided network for human parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4473–4482.
20. Li, P.; Xu, Y.; Wei, Y.; Yang, Y. Self-correction for human parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 3260–3271. [CrossRef] [PubMed]
21. Ihsan, A.M.; Loo, C.K.; Naji, S.A.; Seera, M. Superpixels features extractor network (SP-FEN) for clothing parsing enhancement. *Neural Processing Letters.* **2020**, *51*, 2245–2263. [CrossRef]
22. Ruan, T.; Liu, T.; Huang, Z.; Wei, Y.; Wei, S.; Zhao, Y. Devil in the details: Towards accurate single and multiple human parsing. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 4814–4821.
23. Park, S.; Nie, B.X.; Zhu, S.-C. Attribute and-or grammar for joint parsing of human pose, parts and attributes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 1555–1569. [CrossRef] [PubMed]
24. Gong, K.; Gao, Y.; Liang, X.; Shen, X.; Wang, M.; Lin, L. Graphonomy: Universal human parsing via graph transfer learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7450–7459.
25. Xia, F.; Zhu, J.; Wang, P.; Yuille, A. Pose-guided human parsing by an and/or graph using pose-context features. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
26. Zhang, Z.; Su, C.; Zheng, L.; Xie, X. Correlating edge, pose with parsing. In Proceedings of the IEEE/CVF Conference on computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8900–8909.
27. Chu, J.; Jin, L.; Fan, X.; Teng, Y.; Wei, Y.; Fang, Y.; Xing, J.; Zhao, J. Single-Stage Multi-human Parsing via Point Sets and Center-based Offsets. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 1863–1873.

28. Dou, S.; Jiang, X.; Tu, Y.; Gao, J.; Qu, Z.; Zhao, Q.; Zhao, C.J.a.p.a. DROP: Decouple Re-Identification and Human Parsing with Task-specific Features for Occluded Person Re-identification. *arXiv* **2024**, arXiv:2401.18032.
29. He, R.; Cheng, M.; Xiong, M.; Qin, X.; Liu, J.; Hu, X. Triple attention network for clothing parsing. In Proceedings of the International Conference on Neural Information Processing, Bangkok, Thailand, 18–22 November 2020; pp. 580–591.
30. Guo, H.; Xie, F.; Soong, F.; Wu, X.; Meng, H. A multistage multi-codebook vq-vae approach to high-performance neural tts. *arXiv* **2022**, arXiv:2209.10887.
31. Liu, S.; Feng, J.; Domokos, C.; Xu, H.; Huang, J.; Hu, Z.; Yan, S. Fashion parsing with weak color-category labels. *IEEE Trans. Multimed.* **2013**, *16*, 253–265. [CrossRef]
32. Yamaguchi, K.; Hadi Kiapour, M.; Berg, T.L. Paper doll parsing: Retrieving similar styles to parse clothing items. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3519–3526.
33. Park, T.; Liu, M.-Y.; Wang, T.-C.; Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 2337–2346.
34. Han, X.; Wu, Z.; Huang, W.; Scott, M.R.; Davis, L.S. Finet: Compatible and diverse fashion image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4481–4491.
35. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An empirical study of spatial attention mechanisms in deep networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6688–6697.
36. Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; Tang, X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1096–1104.
37. Ji, W.; Li, X.; Zhuang, Y.; Bourahla, O.E.F.; Ji, Y.; Li, S.; Cui, J. Semantic Locality-Aware Deformable Network for Clothing Segmentation. In Proceedings of the IJCAI, Stockholm, Sweden, 13–19 July 2018; pp. 764–770.
38. Martinsson, J.; Mogren, O. Semantic segmentation of fashion images using feature pyramid networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019; pp. 0-0.Liang, X.; Liu, S.; Shen, X.; Yang, J.; Liu, L.; Dong, J.; Lin, L.; Yan, S. Deep human parsing with active template regression. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 2402–2414.
39. Liang, X.; Liu, S.; Shen, X.; Yang, J.; Liu, L.; Dong, J.; Lin, L.; Yan, S. Deep human parsing withactive template regression. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 2402–2414. [CrossRef] [PubMed]
40. Yang, L.; Rodriguez, H.; Crucianu, M.; Ferecatu, M. Fully convolutional network with superpixel parsing for fashion web image segmentation. In Proceedings of the MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, 4–6 January 2017; pp. 139–151.
41. Xue, H.; Liu, C.; Wan, F.; Jiao, J.; Ji, X.; Ye, Q. Danet: Divergent activation for weakly supervised object localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6589–6598.
42. Yurtkulu, S.C.; Şahin, Y.H.; Unal, G. Semantic segmentation with extended DeepLabv3 architecture. In Proceedings of the 2019 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, 24–26 April 2019; pp. 1–4.
43. Zhou, J.; Hao, M.; Zhang, D.; Zou, P.; Zhang, W. Fusion PSPnet image segmentation based method for multi-focus image fusion. *IEEE Photon-J.* **2019**, *11*, 6501412. [CrossRef]
44. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 603–612.
45. Su, Z.; Chen, M.; Huang, E.; Lin, G.; Zhou, F.J.N. MVSN: A multi-view stack network for human parsing. *Neurocomputing* **2021**, *465*, 437–450. [CrossRef]