Enhanced Context Learning with Transformer for Human Parsing

Jingya Song 1,2,3, Qingxuan Shi 1,2,3,*, Yihang Li 1,2,3 and Fang Yang 1,2,3

1 School of Cyber Security and Computer, Hebei University, Baoding 071002, China
2 Hebei Machine Vision Engineering Research Center, Hebei University, Baoding 071002, China
3 Institute of Intelligent Image and Document Information Processing, Hebei University, Baoding 071002, China

* Correspondence: qingxuanshi@hbu.edu.cn

Abstract: Human parsing is a fine-grained human semantic segmentation task in the field of computer vision. Due to the challenges of occlusion, diverse poses and a similar appearance of different body parts and clothing, human parsing requires more attention to learn context information. Based on this observation, we enhance the learning of global and local information to obtain more accurate human parsing results. In this paper, we introduce a Global Transformer Module (GTM) via a self-attention mechanism to capture long-range dependencies for effectively extracting context information. Moreover, we design a Detailed Feature Enhancement (DFE) architecture to exploit spatial semantics for small targets. The low-level visual features from CNN intermediate layers are enhanced by using channel and spatial attention. In addition, we adopt an edge detection module to refine the prediction. We conducted extensive experiments on three datasets (i.e., LIP, ATR, and Fashion Clothing) to show the effectiveness of our method, which achieves 54.55% mIoU on the LIP dataset, 80.26% on the average F-1 score on the ATR dataset and 55.19% on the average F-1 score on the Fashion Clothing dataset.

Keywords: human parsing; semantic segmentation; deep learning

1. Introduction

Human parsing, a special semantic segmentation task, aims to segment the human body into multiple semantic parts at the pixel level. It plays a potential role in many vision applications, such as video surveillance [1], autonomous driving [2], person re-identification [3], human-computer interaction [4], and fashion synthesis [5].

Recent studies developed several solutions from different perspectives to boost the performance of this task. Previous work with CE2P [6] proved that context information and high-resolution maintenance are two key factors in human parsing solutions. Existing human parsing methods [7–10] fuse multi-scale features to obtain context information. For instance, PSPNet [11] uses the Pyramid Pooling Module (PPM) to capture context information by applying average pooling operations. Although they enlarge the receptive field to aggregate multi-scale features, they fail to learn richer global context dependencies from a global perspective. In addition, many methods maintain high-resolution by directly manipulating the convolution operation to recover missing local information such as linear interpolation and transposed convolution. Chen et al. [12] incorporated details of low-level feature maps.

Previous work on CE2P took advantage of the PPM to obtain context information and utilized a high-resolution module to recover details. In our work, we observe that richer context information and enhanced detail information can be further fully captured. As shown in Figure 1, several unreliable prediction examples generated by CE2P reflect some drawbacks. When the labeled annotations in the human parsing dataset are imbalanced, pixels that do not belong to the human body are prone to misjudgment. This situation
is shown in Figure 1a. Roughly more than half of the pixels are labeled as background, resulting in inaccurate predictions of body parts and clothing. Besides, it is difficult to distinguish symmetrical human body parts and a visual similarity of appearance. The left and right feet, left and right legs have similar appearances as well as similar positions but belong to different categories. For example, as shown in Figure 1b, the coat and upper-clothes have similar appearances that can easily confuse labels with different semantics. Meanwhile, human pose variations and object occlusion may bring out wrong results, as is illustrated in Figure 1c, where the pixels belonging to the human body are missed. A simple $1 \times 1$ convolution is used to obtain low-dimensional features in CE2P, which is insufficient to learn low-level feature maps. This operation may cause failures in segmenting some small targets. Most of them are on the face or the neck, not on the legs; our goal is to enhance the detailed features.

**Figure 1.** Illustration of unreliable human parsing results in CE2P [6]. (a) Inaccurate human body area recognition; (b) Similar appearances are confused; (c) Unidentified human information.

To obtain more accurate parsing results, we propose an enhanced context learning with Transformer for human parsing, which can improve the accuracy in the current field of human parsing both locally and globally. Different from the traditional CNN method, we introduce a Global Transformer Module (GTM) to leverage self-attention to...
capture the global context. To compensate for the loss of detailed features, our method employs a Detailed Feature Enhancement (DFE) architecture to exploit low-level spatial information from CNN features. Specifically, the DFE model includes a channel attention module and a spatial attention module. Channel attention selectively focuses on a featured map, and the combination of the two modules can capture more comprehensive detailed information. We design this module without down-sampling to retain sufficient local structures for small targets. Therefore, small targets such as the glove, scarf and hat can be clearly identified. Considering that human parsing is a fine-grained segmentation task, edge contour information is crucial to support parsing predictions. Experimental results suggest that our architecture presents a better way to leverage self-attention compared with previous CNN-based encoder-decoder methods.

The main contributions of this paper can be summarized as follows:

(1) We propose an enhanced context learning with Transformer for the human parsing method, which can improve the accuracy in the current field of human parsing both locally and globally.

(2) We design a GTM architecture to explore long-range global information through the self-attention mechanism.

(3) To capture fine-grained local information effectively, we design the DFE module to integrate information between the GTM and edge detection module, which learns rich and discriminative detailed features.

2. Related Work

2.1. Human Parsing

Recently, many deep learning methods have been devoted to human parsing. Liu et al. [6] designed a context embedded network (CE2P) based on edge perception by containing three modules of feature resolution, context information and edge detection to achieve a complementary human body analysis. Liang et al. [7] proposed a Co-CNN framework, which can simultaneously capture both local and global information for the human body analysis. Gong et al. [13] proposed a richer and more diverse dataset named Look into Person (LIP), which combined pose estimation with human parsing to obtain richer semantic information and improve parsing results. Nie et al. [14] used the MuLA network to solve the problems in human parsing and pose estimation in parallel, and adjusted effective information by learning from each other to obtain more accurate results. In order to obtain more accurate results, pose estimation or edge detection information is commonly used and can better understand human semantics. Chen et al. [15] introduced the edge-aware filtering method to obtain semantic contour information of adjacent parts, which greatly improved computational efficiency compared with traditional methods. Gong et al. [16] adopted a part grouping network (PGN) to share intermediate features to achieve semantic part segmentation and edge detection. In view of the problem of label confusion in semantic segmentation tasks, Li et al. [17] introduced a self-correcting process method (SCHP) that can eliminate label noises such as an inaccurate boundary, chaotic fine-grained classification and multi-person occlusion, thus improving the reliability of labels and models. Zhang et al. [18] proposed a correlation parsing machine (CorrPM) based on the combination of three modules of pose estimation, human body analysis and edge detection, which utilized the human body key points to correspond the segmented categories to the body parts to which they belonged, and obtained accurate parsing results. CE2P demonstrated that context information, high-resolution and edge information are beneficial for human parsing tasks. However, these methods do not have sufficient access to global and local information, limiting the ability to capture the distribution of different categories and leading to limited performance for human parsing with fine-grained categories. To mitigate this issue, we propose the GTM and DFE to obtain global and local information to enhance context learning for better human parsing performance.
2.2. Context Information Extraction

In view of the semantic segmentation task, exploring the context information can make semantic segmentation results more accurate. ASPP [9] utilized the convolutions of different receptive fields to capture context information. DenseASPP [10] is further improved on the basis of the ASPP network, which introduced dense connections into ASPP to generate multi-scale features to retain more semantic information. Zhao et al. [11] proposed a PSPNet network that integrated multi-scale features with a pyramid pooling module to realize the capture of global context information. ParseNet [19] used global pooling to compute global features and enhanced the feature representation of each pixel. He et al. [20] proposed the Adaptive Pyramid Context network (APCNet) for semantic segmentation, which is composed of multiple ACM blocks; each ACM calculates the context vector of each region by using the local affinity of global guidance. OCRNet [21] transforms the pixel classification problem into an object area classification problem, mainly extracting rich semantic context information by enhancing object information. Despite their success, a limitation of these networks is that they do not perform well in learning the global context and long-range spatial dependencies, which may lead to inconsistent segmentation inside large objects or inferior results for small categories. Thus, in this paper, we propose the GTM to obtain global context information via the self-attention mechanism.

2.3. Transformer in Vision

Transformer was originally used in machine translation and has been widely adopted in Natural Language Processing (NLP) [22,23]. Vaswani et al. [24] employed self-attention instead of traditional RNN to compare sequences in pairs directly to obtain global information further and solve the long-distance dependence problem. Jacob et al. [25] proposed a new language model BERT to train Transformer in an unlabeled text by preprocessing the left and right context. In addition, Transformer is also adopted in chemistry, life sciences, audio processing and even other disciplines. With the development of deep learning, Transformer models have become popular for computer vision tasks and can achieve better results than ordinary neural networks. Parmar et al. [26] used the self-attention mechanism to query each pixel in locally adjacent areas. Transformer and CNN can also be combined due to the similarity of the self-attention mechanism and convolution layer. DETR [27] applied Transformer to the computer vision task and inputted the image features extracted from CNN into the Transformer structure through a series of encoding and decoding to obtain the result of target detection. Deformable DETR [28] added deformable convolutions on the basis of DETR, which can process information in space more efficiently. ViT [29] achieved good performance in the image classification task. Zheng et al. [30] used the ViT structure to extract features and the decoder to restore resolution. Feature maps extracted by SETR are single and low-resolution. SegFormer [31] adopted ViT of the pyramid structure to obtain multi-scale features, which reduces the amount of calculation. Inspired by the success of the self-attention mechanism in these tasks, we apply this mechanism in human parsing. Unlike most transformer models, we only use the encoder to capture global information.

3. Method

The pipeline of our proposed method is shown in Figure 2. Specifically, we adopt ResNet-101 as the backbone to extract features. Our framework consists of three components; the GTM (Global Transformer Module) is implemented upon the output of concatenating the 4-level pyramid to capture rich context information. The DFE (Detailed Feature Enhancement Module) enhances detailed features from conv2 and integrates with the output of GTM to obtain a coarse prediction. In addition, the Edge Detection module uses the learned contour representation to refine the coarse prediction to generate the final human parsing prediction.
3.1. Global Transformer Module

The human body has a semantic structure context prior to be analysed. Thus, capturing the context information from a global perspective can effectively reduce parsing prediction errors, especially for occlusion, and similar appearance categories. Some previous methods implicitly extracted the structural information by deepening the convolutional layers, but it is difficult to obtain richer global context information. Based on the previous work with CE2P, the PPM pools the shared features extracted from the fifth layer in the residual network and generates $1 \times 1, 2 \times 2, 3 \times 3, 6 \times 6$ multiple scale context features. To exploit the power of the semantic structure context fully, we introduce the GTM (Figure 3) which integrates global context information from the pyramid pooling module. The context features are upsampled by bilinear interpolation to keep the same size as the original feature map. Then, fusion features obtained by $1 \times 1$ convolution reduction channels are input into the GTM as feature sequences. One encoder passes the input feature sequences to the self-attention layer and feed-forward networks for processing, and the output result is passed to the next encoder. Unlike the standard transformer, which adopts the original six layers structure, our encoder is composed of four encoder layers with the same structure. Each encoder has two sub-layers for each layer, and residual connections are employed around each sub-layer, followed by layer normalization. These indicate that the GTM is able to understand the global information due to the self-attention mechanism on the feature sequential information.
3.1.1. Self-Attention Mechanism

As a central piece of GTM, self-attention comes with a flexible mechanism to deal with variable-length inputs. It can be understood as a fully connected layer where the weights are dynamically generated from pairwise relations from input feature sequences. The input vector is first transformed into three different vectors: the query vector $q$, the key vector $k$, and the value vector $v$, and their dimensions are 512. Vectors derived from different inputs are then packed together into three different matrices, namely, $Q$, $K$ and $V$ in Step 1. We calculate the attention scores between each pair of different vectors in Step 2, and these scores determine the degree of attention that we give other features when encoding the human body features at the current position. Step 3 normalizes the scores to enhance gradient stability for improved training, and Step 4 translates the scores into probabilities. Finally, each key vector is multiplied with the softmax score value and then weighted to obtain the value of the current body part node. The detailed algorithm of the self-attention mechanism is summarized in Algorithm 1.

The process of scaled dot product attention (Figure 4a) used by Transformer is given by,

$$ Attention(Q, K, V) = \text{softmax} \left( \frac{Q \cdot K^T}{\sqrt{d_k}} \right) \cdot V. $$

![Figure 3. The architecture of GTM (Global Transformer Module).](image)

![Figure 4. Structure of Self-Attention Mechanism. (a) Scaled Dot-Product Attention; (b) Multi-Head Attention.](image)
3.1.2. Positional Encoding

Since there is no convolution in the Transformer, positional encodings are added to the input embedding information at the bottoms of the encoder in order to capture the sequential information. Specifically, the positional encodings used in this paper is the sine type that can be expressed as:

$$PE_{(pos,2i)} = \sin(\text{pos} / 10000^{2i/d_{\text{model}}}),$$

where \text{pos} refers to the position; \text{i} denotes the index number of each value in the vector; and \text{d_{model}} is positional encoding with a dimension of 512.

3.2. Detailed Feature Enhancement Module

In human parsing, we need to classify small targets such as the scarf, sunglasses, socks and glove. Thus, it is essential to explore fine-grained features for pixel-level predictions. Considering that average pooling and consecutive convolution strides’ operations in conventional CNNs, the feature maps shrink during forward propagation. This may make the detailed structures blurred. In order to compensate for the lost detail information, we add a Detailed Feature Enhancement (DFE) module after conv2 of Resnet as a detailed branch. As shown in Figure 5, given the feature map extracted by conv2, our module sequentially infers attention maps along two separate dimensions, channel and spatial; then, the attention maps are multiplied to the input feature map for adaptive feature refinement. Finally, we conduct two sequential \(1 \times 1\) convolutions on the concatenated feature to fuse the local and global context information. The output of features passes through another \(1 \times 1\) convolution to generate the coarse parsing result.

Figure 5. Structure of DFE module.

3.3. Edge Detection Module

The aim of the edge detection module is to learn the representation of the contour to provide assistance for further refine predictions. Hence, we adopt the edge detection module to fuse low-level features with global information and the high-resolution features
to improve human contours’ accuracy. As shown in Figure 2, after the $1 \times 1$ convolution of conv2, conv3, conv4 and the $3 \times 3$ convolution operation, they are upsampled to the same size as conv2 by linear interpolation, and they are fused to obtain feature maps with edge information after the $1 \times 1$ convolution. Finally, these intermediate features from the GTM module, DFE module and edge detection are further concatenated by the $1 \times 1$ convolution to generate the final parsing results. Two cases demonstrate the effect of the edge detection module in Figure 6. In the first case in Figure 6a, from the coarse prediction, the right arm is not clearly identified due to the occlusion of the object on the arm; the socks on the left and right feet belong to the small target object and are prone to misjudgment as the leg. These categories can be clearly distinguished from the final prediction. The second case in Figure 6b shows that some pixels in the right leg are predicted as the left leg, while there is no semantic boundary in this region. The shoe region loses many details in the down-sampling process and cannot be correctly classified. After associating the edge detection module with the parsed features, the model has the awareness of the location of the foot and shoe from the final prediction. Thus, edge features are useful to assist in human parsing.

Figure 6. The effect of edge detection module. (a) The effect of the edge detection module on small targets like sock; (b) The effect of the edge detection module on the categories of left and right symmetrical parts of the human body, avoiding the confusion of categories such as left and right legs and shoes.

3.4. Loss Function

The outputs of our network consist of three components, the coarse parsing result, final parsing result and edge prediction. The total loss can be formulated as:

$$L = L_{\text{coarse-par}} + L_{\text{edge}} + L_{\text{final-par}} + \lambda L_{\text{ar}}$$  \hspace{1cm} (3)
where \( L_{\text{coarse-par}} \) denotes the loss implemented on the branch of parsing with the GTM module and the DFE module. \( L_{\text{edge}} \) is the loss of the edge detection branch. \( L_{\text{final-par}} \) represents the loss of the fusion of the edge detection, GTM module and DFE module. \( L_a \) is the auxiliary loss, which implements on the intermediate feature map output from the conv4 block.

Every loss in Equation (3) using cross-entropy loss can be defined as follows:

\[
L_s = \sum_{i=1}^{H \times W} \sum_{n=1}^{N} -y_{in} \log \hat{y}_{in},
\]

where \( N \) denotes the total number of categories; \( \hat{y}_{in} \) denotes the predicted probability value that the \( i \)-th pixel belongs to category \( n \); \( y_{in} \) denotes the true probability that the \( i \)-th pixel belongs to category \( n \). \( H \times W \) represents the total number of pixels.

4. Experiments

4.1. Datasets and Metrics

Datasets. We demonstrate the performance of our methods on three human parsing datasets, including the Look Into Person (LIP) dataset [13], Fashion Clothing Dataset [32] and Active Template Regression (ATR) [33].

The LIP dataset is the largest single human dataset that provides 50,462 images with 19 semantic human part labels and 16 body key points. The 19 semantic human part labels contain the hat, hair, glove, sunglasses, upper-clothes, dress, coat, socks, pants, jumpsuits, scarf, skirt, face, left-arm, right-arm, left-leg, right-leg, left-shoe, and right-shoe. The images in the LIP were collected from realistic scenes with challenging poses, heavy occlusions and complex backgrounds. Images in this dataset are divided into 30,462 for training, 10,000 for validation and 10,000 for testing.

The Fashion Clothing Dataset is a collection of 4371 images from Colorful Fashion Parsing [34], Fashionista [35] and Clothing Co-Parsing [36]. It is split into 1716 and 1817 images for training and testing. One background and 17 pixel-level labels are annotated. A total of 17 pixel-level labels mainly focus on clothing details, i.e., jewelry, bags, coats, belts, dresses, glasses, hair, pants, shoes, shirts, skin, skirts, upper-clothes, vests and underwear, scarves, socks and hats. Following the label set defined by Dong et al. [34], we merge the labels of the Fashionista and CFPD datasets into 18 categories: faces, sunglasses, hats, scarves, hair, upper-clothes, left-arm, right-arm, belts, pants, left-leg, right-leg, skirts, left-shoe, right-shoe, bags, dresses and background.

The ATR dataset is the first large dataset that appeared in the field of human parsing. There are 6000 images for training, 700 images for validation and 1000 images for testing. The ATR contains 18 categories including background, hats, hair, sunglasses, upper-clothes, skirts, pants, dresses, belts, left/right-shoes, face, left/right-legs, left/right-arms, bags and scarves.

Metrics. For the LIP dataset, we follow three metrics to evaluate human parsing results, pixel-wise accuracy (Pixel Acc.), mean accuracy (Mean Acc.) and mean pixel Intersection-over-Union (mIoU). The pixel accuracy, foreground accuracy (F.G. Acc.), average precision (Avg. P.), average recall (Avg. R.) and average F-1 are leveraged as the evaluation metrics for the ATR and Fashion Clothing datasets.

4.2. Implement Details

During training, the input image size is \( 384 \times 384 \). We use a “Poly” learning rate policy for a total of 150 epochs with a base learning rate of 0.001. The momentum and weight are set as 0.9, 0.0005, respectively. For data augmentation, we apply the random scaling (from 0.5 to 1.5) and left-right flipping during training. We adopt cross-entropy loss when training on all datasets. The weight parameter \( \lambda \) in Equation (3) is set as 0.4.
4.3. Quantitative Results

**LIP.** We compare the performance of our network with other state-of-the-art methods on the LIP dataset. As shown in Table 1, the proposed method yields the result of 54.5% in terms of mIoU. Compared with the CE2P method, our method exceeds by 1.1%, 2.89% and 0.29% in terms of mIoU, Mean Acc. and Pixel Acc., respectively. In order to verify the detailed effectiveness of our structure, we further present the per-class IoU in Table 2. While comparing with the current state-of-the-art method CE2P, the DFE module significantly improves the performance of the small classes such as the scarf, glove and hat, which demonstrates its ability to capture low-level features. Our proposed approach achieves a large gain especially for some confusing categories such as upper-clothes, coats, dresses and skirts. From the viewpoint of the IoU of the dress and skirt, our method yields obviously an improvement approximately 5% and 6% better than CE2P. These improvements imply that the GTM module generates more global information via self-attention. Moreover, edge detection in our network utilizes edge features to assist human parsing. The left and right arms have a similar appearance; edge detection can identify them clearly. These superior performance experimental results further show that our method is capable of enforcing feature information globally and locally.

![Table 1. Comparison of different methods on the validation set of the LIP dataset.](image)

| Method                  | Pixel Acc. | Mean Acc. | mIoU  |
|-------------------------|------------|-----------|-------|
| DeepLab [9]             | 82.66      | 51.64     | 41.64 |
| Attention [37]          | 83.43      | 54.39     | 42.92 |
| DeepLab (ResNet-101)    | 84.09      | 55.62     | 44.80 |
| MuLA [14]               | 88.50      | 60.50     | 49.30 |
| JPPNet [13]             | 86.39      | 62.32     | 51.37 |
| CE2P [6]                | 87.37      | 63.20     | 53.10 |
| Ours                    | 87.66      | 66.09     | 54.50 |

**Fashion Clothing Dataset.** Table 3 reports the results and comparisons with three recent approaches on the Fashion Clothing dataset. Our method outperforms the DeepLab by 19.05% and 17.01%, in terms of average recall and average precision scores, respectively. Compared with the Attention method, our method significantly improves the pixel accuracy, foreground accuracy, average accuracy, average recall and average F-1 score and achieves 92%, 66.16%, 54.4%, 56.01% and 55.19%, respectively. This performance suggests the superiority of our parsing method with the assistance of detailed feature and edge factors, which can introduce contextual cues into the human parsing task.

**ATR.** The evaluation results for the test set of the ATR dataset are given in Table 4. The proposed method has a significant performance improvement in most of the metrics. It is 3.23% and 0.45% higher than Co-CNN in terms of average recall and average precision. It confirms the effectiveness of the Transformer self-attention mechanism for human parsing and illustrates that the GTM module has a strong capability to incorporate global and local information.

4.4. Qualitative Comparison

We provide the quality results on the LIP dataset in Figure 7. Compared with CE2P, our results are more reasonable. For example, in Figure 7b, the coat and upper-clothes are easily confused. There are inaccurate boundaries between adjacent parts of the body resolution and many easily confused categories, such as the coat and upper-clothes in the second column; in the fifth column, long skirts with similar appearances to short skirts are easily misidentified. Benefitting from the GTM, our method can correctly predict them. With the help of DFE, we also observe from the last row that the model can learn the detail information and accurately identifies the left/right arm and areas obscured by objects. Consequently, our method can obtain reasonable and precise results.
Table 2. Performance comparison in terms of mean pixel Intersection-over-Union (mIoU) with state-of-the-art methods on LIP validation set.

| Method       | Bkg | Hat | Hair | Glove | Glass | u-Cloth | Dress | Coat | Socks | Pants | j-Suits | Scarf | Skirt | Face | l-Arm | r-Arm | l-Leg | r-Leg | l-Shoe | r-Shoe | mIoU  |
|--------------|-----|-----|------|-------|-------|---------|-------|------|-------|-------|---------|-------|-------|------|-------|-------|-------|-------|-------|-------|-------|-------|
| SegNet [5]   | 70.62 | 26.60 | 44.01 | 0.01  | 0.00  | 34.46   | 0.00  | 15.97| 3.59  | 33.56 | 0.01    | 0.00  | 0.00  | 52.38| 15.30 | 24.23 | 13.82 | 9.26  | 6.47  | 18.17 |
| DeepLab [9]  | 83.25 | 57.94 | 66.11 | 28.50 | 18.40 | 60.94   | 23.17 | 47.03| 34.51 | 64.00 | 22.38   | 14.29 | 18.74 | 69.70| 49.44 | 51.66 | 37.49 | 34.60 | 28.22 | 22.41 | 41.64 |
| Attention [37]| 84.00 | 58.87 | 66.78 | 23.32 | 19.48 | 63.20   | 29.63 | 49.70| 35.23 | 66.04 | 24.73   | 12.84 | 20.41 | 70.58| 50.17 | 54.03 | 38.35 | 37.70 | 26.20 | 27.09 | 42.92 |
| DeepLab101   | 84.09 | 59.76 | 66.22 | 28.76 | 23.91 | 64.95   | 33.68 | 52.86| 37.67 | 68.05 | 26.15   | 17.44 | 25.23 | 70.00| 50.42 | 53.89 | 39.36 | 38.27 | 26.95 | 28.36 | 44.80 |
| JPPNet [13]  | 86.26 | 63.55 | 70.20 | 36.16 | 23.48 | 68.15   | 31.42 | 55.65| 44.56 | 72.19 | 28.39   | 18.76 | 25.14 | 73.36| 61.97 | 63.88 | 58.21 | 57.99 | 44.02 | 44.09 | 51.37 |
| CE2P [6]     | 87.67 | 65.29 | 72.54 | 39.09 | 32.73 | 69.46   | 32.52 | 56.28| 49.67 | 74.11 | 27.23   | 14.19 | 22.51 | 75.50| 65.14 | 66.59 | 60.10 | 58.59 | 46.63 | 46.12 | 53.10 |
| Ours         | 87.87 | 67.28 | 72.11 | 42.24 | 33.70 | 70.46   | 37.82 | 57.07| 50.55 | 75.21 | 32.16   | 17.67 | 28.73 | 74.84| 65.54 | 67.83 | 59.60 | 58.81 | 45.28 | 45.42 | 54.50 |
Table 3. Comparison of different methods on the test of Fashion Clothing.

| Method        | Pixel Acc. | E.G.Acc | Avg.P. | Avg.R. | Avg.F-1 |
|---------------|------------|---------|--------|--------|---------|
| DeepLab [9]   | 87.68      | 56.08   | 35.35  | 39.00  | 37.09   |
| Attention [37]| 90.58      | 64.47   | 47.11  | 50.35  | 48.68   |
| Ours          | 92.00      | 66.16   | 54.40  | 56.01  | 55.19   |

Table 4. Comparison of different methods on the ATR test dataset.

| Method        | Pixel Acc. | E.G.Acc | Avg.P. | Avg.R. | Avg.F-1 |
|---------------|------------|---------|--------|--------|---------|
| ATR [33]      | 91.11      | 71.04   | 71.69  | 60.25  | 64.38   |
| DeepLab [9]   | 94.42      | 82.93   | 78.48  | 69.24  | 73.53   |
| PSPNet [11]   | 95.20      | 80.23   | 79.66  | 73.79  | 75.84   |
| Attention [37]| 95.41      | 85.71   | 81.30  | 73.55  | 77.23   |
| Co-CNN [7]    | 96.02      | 83.57   | 84.95  | 77.66  | 80.14   |
| Ours          | 96.10      | 84.02   | 84.87  | 80.89  | 80.26   |

Figure 7. Visualization of different methods on the LIP validation dataset. (a) Comparison with CE2P method in terms of object occlusion; (b) In terms of comparing similar appearance categories, the CE2P method tends to confuse coat with upper-clothes; (c) Compared with multi-person in crowded scenes, our method is more accurate; (d) In terms of comparing left and right symmetrical body parts, our method can clearly identify left and right arms; (e) In the comparison of categories with similar appearance, the CE2P method misclassifies dress as skirt; (f) Left and right arms are similar in appearance and easily misjudged in CE2P method.
4.5. Ablation Study

We perform extensive ablation experiments to illustrate the effect of each component of our method; the result is shown in Table 5.

4.5.1. The Effect of GTM

In order to evaluate the effectiveness of each module, we first introduce the GTM module for experiments, which can generate global information to help us obtain more accurate results. As shown in Table 5, without the GTM, DFE and edge detection are denoted as a baseline model B. The baseline model achieves 51.54% mIoU. Adding GTM to the baseline model can be named as B + T; the result shows a gain of mIoU by 2.01 points. It indicates that the self-attention mechanism in the GTM module can acquire global information via long-range dependencies to assist the human parsing task. Compared with baseline model B, the performances on some classes which are usually adjacent and have easily confused appearances (e.g., upper-clothes and j-suits) gain nearly 1% and 6% mIoU, respectively. The dress and skirt boost by 5.26% and 7.25% in terms of mIoU. As can be seen, the performance on some small targets of the scarf and glove reach 20.07% and 40.08%, respectively. Since the long-range semantic information can provide more discriminated features, the GTM model is beneficial for recognizing large-size objects.

We set multiple layers of encoder in the GTM. To explore the optimal encoder layers, we design four variants with different encoder layers. The best performance when encoder layers are set to 4, we find that more encoder layers do not yield better performance.

4.5.2. The Effect of DFE

Since human parsing is a fine-grained semantic segmentation task, rich detailed semantic information is essential to identify small targets. Using the high-resolution module in CE2P can recover details, but we argue it may obtain more effective detailed features. Thus, we introduce the DFE module to grasp the detailed features. In Table 5, when there are no T and E involved, our method has a significant improvement in object recognition for small targets such as gloves and scarves compared to baseline. By employing the T, D at the same time, we can see that the categories of large-size objects, dresses and skirts, and small target socks have improved in performance. It shows that the GTM and DFE modules in our method mutually promote each other. We further conduct ablation experiments on each part of DFE, which consists of spatial attention (SA) and channel attention (CA). As shown in Table 6, when we added the CA module, the performance on the class of the skirt and scarf improve 1.02% and 0.99%, respectively. Meanwhile, the left and right symmetrical body parts, such as the left and right arms, are also enhanced. With the spatial and channel feature derived from the DFE, more details that are not available in deep layers are provided for parsing from shallow and high-resolution layers. The experimental results also demonstrate the effectiveness of our DFE module.

4.5.3. The Effect of Edge Detection Module

The edge detection module plays a role in guiding the prediction of parsing. It can separate the body part from contour information. The model B + T + E denotes concatenating both the GTM and edge detection, which yields about 2.3% improvement in terms of mIoU compared to the baseline model. This gain is mainly due to the accurate prediction at the boundary area between semantic parts. By introducing the edge information, the improvement is close to 3% mIoU for some categories with similar or adjacent appearances, such as socks and pants to the experimental data in Table 5. These results highlight that the edge detection module can further optimize the challenge of prediction accuracy especially in some small regions and unclear edge contours categories.
Table 5. Comparison of per-class IoU on the LIP validation set. “B” means baseline module; “T” means GTM; “E”, “D” denote edge detection module and DFE module.

| Method          | Bkg | Hat | Hair | Glove | Glass | u-Cloth | Dress | Coat | Socks | Pants | j-Suits | Scarf | Skirt | Face | l-Arm | r-Arm | l-Leg | r-Leg | l-Shoe | r-Shoe | mIoU |
|-----------------|-----|-----|------|-------|-------|---------|-------|------|-------|-------|---------|-------|-------|------|-------|-------|-------|-------|-------|-------|-------|
| B               | 87.22 | 65.34 | 72.13 | 36.18 | 31.97 | 68.86   | 31.02 | 55.81 | 47.35 | 73.23 | 26.91   | 12.28 | 20.58 | 74.49 | 62.95 | 65.18 | 56.31 | 55.99 | 43.49 | 43.80 | 51.54 |
| B + T           | 87.70 | 65.87 | 71.77 | 40.08 | 29.76 | 69.46   | 36.48 | 56.39 | 46.69 | 74.51 | 33.14   | 20.07 | 27.83 | 74.68 | 64.07 | 67.00 | 57.77 | 57.69 | 44.40 | 45.68 | 53.55 |
| B + D           | 86.72 | 63.72 | 70.41 | 40.00 | 27.87 | 66.84   | 34.03 | 53.00 | 44.98 | 72.37 | 26.33   | 15.56 | 27.37 | 73.54 | 62.48 | 64.80 | 56.82 | 55.96 | 42.88 | 43.65 | 51.47 |
| B + T + D       | 87.63 | 65.68 | 71.77 | 39.90 | 28.95 | 69.19   | 34.66 | 55.48 | 47.31 | 74.73 | 34.22   | 21.06 | 28.40 | 74.38 | 64.19 | 66.77 | 57.90 | 57.58 | 44.48 | 45.20 | 53.47 |
| B + T + E       | 87.86 | 66.84 | 71.72 | 42.72 | 30.81 | 69.95   | 36.62 | 56.90 | 49.46 | 74.97 | 31.50   | 18.76 | 25.26 | 74.81 | 65.26 | 67.37 | 58.34 | 57.95 | 44.45 | 45.26 | 53.84 |
| B + T + E + D   | 87.86 | 67.15 | 71.70 | 43.45 | 32.40 | 70.08   | 38.10 | 56.59 | 50.08 | 74.80 | 31.50   | 17.50 | 27.21 | 74.83 | 64.97 | 67.57 | 58.79 | 58.36 | 44.35 | 45.43 | 54.14 |
| B + T + E + D (auxiliary loss) | 87.87 | 67.28 | 72.11 | 42.24 | 33.70 | 70.46   | 37.82 | 57.07 | 50.55 | 75.21 | 32.16   | 17.67 | 28.73 | 74.84 | 65.54 | 67.83 | 59.60 | 58.81 | 45.28 | 45.42 | 54.50 |

Table 6. Comparison of different components in DFE module.

| Method          | Bkg | Hat | Hair | Glove | Glass | u-Cloth | Dress | Coat | Socks | Pants | j-Suits | Scarf | Skirt | Face | l-Arm | r-Arm | l-Leg | r-Leg | l-Shoe | r-Shoe | mIoU |
|-----------------|-----|-----|------|-------|-------|---------|-------|------|-------|-------|---------|-------|-------|------|-------|-------|-------|-------|-------|-------|-------|
| B + T + SA      | 87.69 | 66.45 | 71.96 | 41.14 | 28.25 | 69.38   | 35.28 | 56.71 | 47.93 | 74.84 | 32.8   | 20.07 | 27.38 | 74.53 | 63.78 | 66.26 | 57.51 | 57.27 | 44.95 | 45.39 | 53.48 |
| B + T + SA + CA | 87.63 | 65.68 | 71.77 | 39.90 | 28.95 | 69.19   | 34.66 | 55.48 | 47.31 | 74.73 | 34.22  | 21.06 | 28.40 | 74.38 | 64.19 | 66.77 | 57.90 | 57.58 | 44.48 | 45.20 | 53.47 |
4.5.4. The Effect of Auxiliary Loss  

On the basis of our method, we further discuss the role of \( L_a \). Auxiliary loss brings an improvement of 0.36% in terms of mean IoU, as shown in Table 5. Therefore, we can infer that the auxiliary loss during the training procedure facilitates the learning of semantic feature maps from the conv4 stage to the conv5 stage. The high-level semantic information obtained from conv5 can further assist in better parsing results. In addition, the GTM model can exploit this information to obtain rich global feature information.

5. Conclusions  

In this paper, we propose the Global Transformer Module (GTM) and Detailed Feature Enhancement (DFE) to improve human parsing performance by learning global and local information. We rethink the human parsing task by considering the context information. In this way, the GTM with self-attention can learn rich global feature representations. With the help of the DFE module, the network has more chance of obtaining low-level features to identify small classes. Moreover, our approach utilizes edge detection to learn contour representation to obtain more accurate prediction further. Extensive experimental results on three datasets show that our method can outperform recent methods. In the future, we plan to add pose estimation for multi-task learning. We hope that pose information as guidance will boost the performance of parsing.

Author Contributions: Methodology, J.S.; software, Y.L.; validation, J.S., Y.L. and Q.S.; formal analysis, J.S.; investigation, Q.S., J.S. and Y.L.; writing—original draft preparation, J.S.; writing—review and editing, J.S., Q.S., Y.L. and F.Y.; visualization, J.S.; supervision, Q.S.; funding acquisition, Q.S. and F.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by The Natural Science Foundation of Hebei Province (F2019201451).  
This work was also supported by Science and Technology Project of Hebei Education Department (ZD2019131).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Lu, Y.; Boukhourba, K.; Boonnaert, J.; Fleury, A.; Lecoeuche, S. Application of an incremental SVM algorithm for online human recognition from video surveillance using texture and color features. Neurocomputing 2014, 126, 132–140. [CrossRef]
2. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
3. Kalayeh, M.M.; Basaran, E.; Gökmen, M.; Kamasak, M.E.; Shah, M. Human semantic parsing for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1062–1071.
4. Qi, S.; Wang, W.; Jia, B.; Shen, J.; Zhu, S.C. Learning human-object interactions by graph parsing neural networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 401–417.
5. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image seg-mentation. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 2481–2495. [CrossRef] [PubMed]
6. 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.
7. Liang, X.; Xu, C.; Shen, X.; Yang, J.; Liu, S.; Tang, J.; Lin, L.; Yan, S.; Sun Yat-sen University; National University of Singapore; et al. Human parsing with contextualized convolutional neural network. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1386–1394.
8. Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 6–10 June 2015; pp. 3431–3440.
9. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 40, 834–848. [CrossRef] [PubMed]
10. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.

11. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.

12. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.

13. 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.

14. Nie, X.; Feng, J.; Yan, S. Mutual learning to adapt for joint human parsing and pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 502–517.

15. Chen, L.C.; Barron, J.T.; Papandreou, G.; Murphy, K.; Yuille, A.L. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4545–4554.

16. Gong, K.; Liang, X.; Li, Y.; Chen, Y.; Yang, M.; Lin, L. Instance-level human parsing via part grouping network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Munich, Germany, 8–14 September 2018; pp. 770–785.

17. Li, P.; Xu, Y.; Wei, Y.; Yang, Y. Self-correction for human parsing. IEEE Trans. Pattern Anal. Mach. Intell. 2020. [CrossRef] [PubMed]

18. 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.

19. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. arXiv 2015, arXiv:1506.04579.

20. He, J.; Deng, Z.; Zhou, L.; Wang, Y.; Qiao, Y. Adaptive pyramid context network for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7519–7528.

21. Yuan, Y.; Chen, X.; Chen, X.; Wang, J. Segmentation transformer: Object-contextual representations for semantic segmentation. arXiv 2019, arXiv:1909.11065.

22. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv 2019, arXiv:1901.02860.

23. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. Adv. Neural Inf. Process. Syst. 2019, 32. [CrossRef]

24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.

25. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018, arXiv:1810.04805.

26. Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm Sweden, 10–15 July 2018; pp. 4055–4064.

27. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 213–229.

28. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. arXiv 2020, arXiv:2010.04159.

29. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.

30. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Zhang, L.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.

31. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 2021, 34, 12077–12090.

32. Luo, X.; Su, Z.; Guo, J.; Zhang, G.; He, X. Trusted guidance pyramid network for human parsing. In Proceedings of the 26th ACM international conference on Multimedia, Seoul, Korea, 22–26 October 2018; pp. 654–662.

33. 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. [CrossRef] [PubMed]

34. 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]