DETR-based Layered Clothing Segmentation and Fine-Grained Attribute Recognition

Hao Tian Yu Cao P. Y. Mok*
The Hong Kong Polytechnic University, Kowloon, Hong Kong
{hao-henry.tian, yu-daniel.cao}@connect.polyu.hk, tracy.mok@polyu.edu.hk

Abstract

Clothing segmentation and fine-grained attribute recognition are challenging tasks at the crossing of computer vision and fashion, which segment the entire ensemble clothing instances as well as recognize detailed attributes of the clothing products from any input human images. Many new models have been developed for the tasks in recent years, nevertheless the segmentation accuracy is less than satisfactory in case of layered clothing or fashion products in different scales. In this paper, a new DEtection TRansformer (DETR) based method is proposed to segment and recognize fine-grained attributes of ensemble clothing instances with high accuracy. In this model, we propose a ***multi-layered attention module*** *by aggregating features of different scales, determining the various scale components of a single instance, and merging them together. We train our model on the Fashionpedia dataset and demonstrate our method surpasses SOTA models in tasks of layered clothing segmentation and fine-grained attribute recognition.*

1. Introduction

Clothing segmentation and attribute recognition are the fundamental pre-tasks in many fashion applications such as outfit matching, fashion recommendation, and virtual try-on. With a growing interest in fashion related AI research, many researchers devoted themselves to this field and presented excellent work [3]. However, existing methods still face drawbacks when performing layered clothing segmentation and fine-grained attribute recognition tasks. Ge *et al.* [5] proposed a Match R-CNN to integrate clothing detection, landmark regression, segmentation, and retrieval into such a multi-task learning framework trained on DeepFashion2 dataset. However, their method is deficient in the case of clothing vague or layered occlusion and unable to handle multiple tasks harmoniously and simultaneously. Jia *et al.* [8] proposed the Attribute-Mask R-CNN to jointly perform instance segmentation and localized attribute recognition on Fashionpedia dataset with the whole ensemble of clothing instances. Nevertheless, the gap between clothing segmentation and attribute recognition still exists, as well as the incomplete and inferior results for fine-grained attributes. To bridge the gap between instance segmentation and attribute recognition, Xu *et al.* [13] presented Fashionformer by building a DETR-based [1] framework trained on Fashionpedia dataset [8] with a Multi-Layer Rendering module for the attribute stream to explore more fine-grained features. However, their method is not sensitive to clothing or accessories with scale differences and results in missing or incomplete clothing segmentation.

Considering that the tasks of segmentation and attribute recognition complement to one another, we design a model based on DETR that is equipped with a ***multi-layered attention module*** for the segmentation stream in the decoder design. Our method achieves higher average precision and better visual results in clothing segmentation, and addresses to the issue of incomplete mask in case of clothing with different scales. Our main *contributions* are summarized as follows: (1) We build a DETR-based model that can perform well on layered clothing segmentation and fine-grained attribute recognition task; (2) We proposed a new multi-layered attention module in our decoder by aggregating features of fashion items with different scales; and (3) Both qualitative and quantitative comparative analyses show that our method outperforms other SOTA methods in terms of layered clothing segmentation and fine-grained attribute recognition performance.

2. Related Work

Layered Clothing Segmentation Many researchers have attempted various approaches to the task of multi-layer occluded clothing segmentation and attribute recognition. Zheng *et al.* [14] achieved clothing segmentation and attribute recognition with Faster RCNN, SSD, and YOLO on their ModaNet dataset. Jia *et al.* [8] proposed a dataset named Fashionpedia and designed a novel Attribute-Mask

*Corresponding author

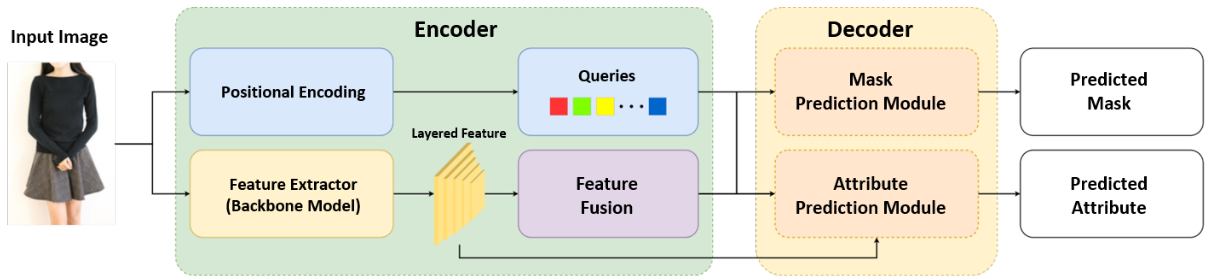


Figure 1. The overall architecture of the DETR-based method with detailed decoder structure illustrated in Fig. 2

R-CNN model to realize the multi-label attribute prediction.

Fashion Datasets In the field of fashion segmentation and attribute recognition, there are currently mainly 4 publicly available datasets, DeepFashion [10], ModaNet [14], Deepfashion2 [5] and Fashionpedia [8]. Fashionpedia is a step toward mapping out the visual aspects of the fashion world, which consists of an ontology built by fashion experts and a dataset with everyday and celebrity event fashion images annotated. Unlike the DeepFashion series, images in Fashionpedia mainly focused on the whole ensemble of clothing. ModaNet [14] focused on outwear and accessories, nevertheless ModaNet server is no longer accessible. We therefore trained the model on the Fashionpedia dataset and conducted comparative experiment.

Transformer-based Detection and Segmentation Carion *et al.* [1] first presented an end-to-end transformer structure to handle object detection task, which also introduced the concept of object query which is a type of input to the transformer decoder. Shi *et al.* [12] used internal attention and external attention for multi-level context mining. *Inspired by these works, a DETR-based transformer architecture is proposed to unify and simplify fashion tasks. The main contribution of this work lies in the decoder design the transformer structure, covering both the segmentation prediction and attribute prediction modules.*

3. Method

3.1. Overall Architecture

Fig. 1 shows the overall architecture of our method, depicting the encoder and decoder parts. In the encoder, we first obtain layered features from the input image by a backbone network with the feature pyramid structure, then fuse the layered feature map into a fused feature and a positional encoder generates positional embeddings at the same time, which we called $Query$, simplified as Q. The layered feature, fused feature, and queries are then sent to the decoder. The decoder consists of a mask prediction module and an attribute prediction module. Similar to [13], the mask prediction module and the attribute prediction module of our model work in a cascaded mode. Instead of generating mask by mask grouping and query learning [13], our method takes fused features and queries as inputs for the mask prediction module, taking into account of both global and local features to improve the mask prediction accuracy, especially for the layered clothing in different scales.

3.2. Encoder

Firstly, we send each input image into the feature extractor to obtain the layered feature map F_i, where $i \in \{1, 2, 3, 4\}$ are indexes of different scales. The feature extractor is a backbone network with a feature pyramid structure, such as Mask R-CNN and Swin Transformer. Then we use a concatenate function to sum up the layered features into a fused feature map F_f. A 1×1 convolution layer is settled to generate a $d \times H \times W$ feature map, which is the query Q. We also obtain the weights of the instance masks in this step, which are also equal to queries weights. The queries are significant input for two prediction modules in the decoder.

3.3. Decoder

We build the decoder with a mask prediction module and an attribute prediction module, as shown in Fig. 2. For the mask prediction module, we first adopt a multi-layered at-

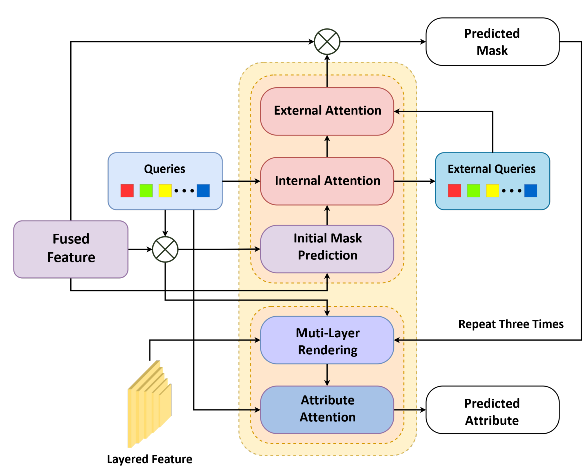
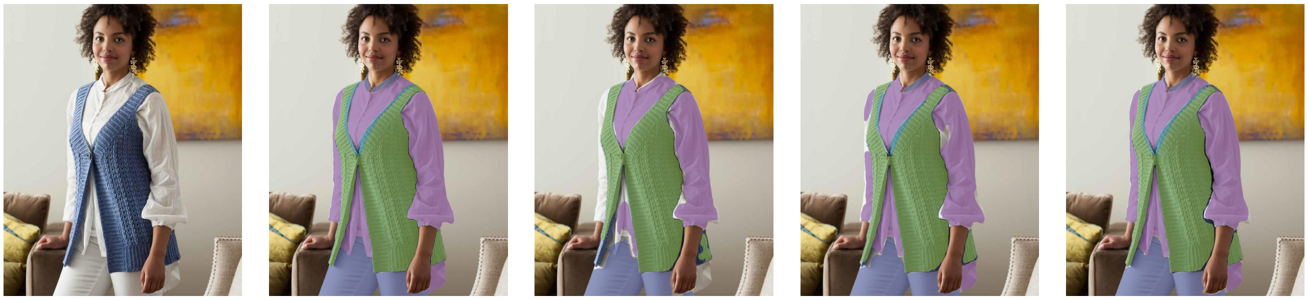
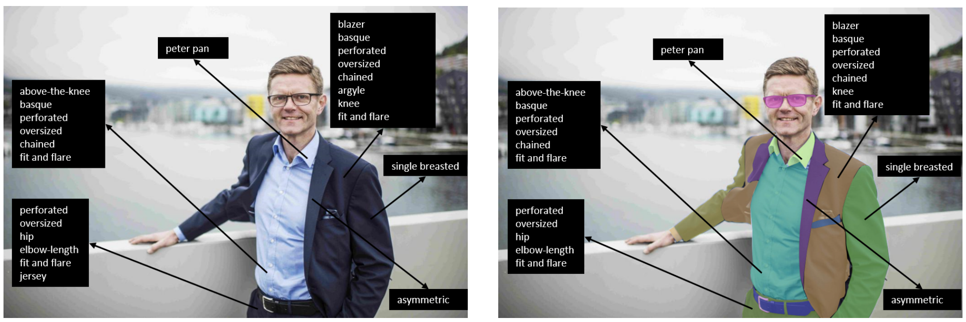


Figure 2. The multi-layered attention module structure in the decoder.



(a) Original Image (b) Groud Truth (c) Attribute-Mask RCNN [8] (d) FashionFormer [13] (e) **Ours**

Figure 3. Experimental result comparison with other SOTA methods.



(a) Input Image and Ground Truth Attributes (b) Our Masks and Attributes Results

Figure 4. An example result of attribute recognition.

tention module with layered features and fused features to obtain the initial mask prediction. We compute the mask prediction by the self-attention in the internal attention and external attention module as the equation below:

$$P_{mask} = SoftMax\left(\frac{QK^T}{\sqrt{C}}\right)F \tag{1}$$

where K is the key positional embeddings, C is the vector dimensions in internal attention and external attention respectively. The fused feature and Queries perform Kronecker product operation to generate the prediction of the initial mask. For internal attention, Q are the queries, K and v come from the initial mask prediction, then the result of this attention can generate a new prediction and a series of new queries which we called $ExternalQueries$. Finally, external attention uses external queries as Q and predictions generated by internal attention to predict next-generation mask prediction, which perform Kronecker product operation with fused feature to generate the final predicted mask.

In the attribute prediction module, we followed the Multi-Layer Rendering (MLR) in [13], so our attribute recognition results are fine-grained as well. The multi-level features are computed as:

$$F_{i,atr}^{j} = \sum_{u}^{W_i}\sum_{v}^{H_i} P^{j-1}(u,v,i)\cdot F_i(u,v)\cdot Q_i^j(u,v) \tag{2}$$

where j is the interaction number, $j \in \{1,2,3\}$, W_i and H_i are the height and width of the corresponding features F_i, u and v are the spatial index of features.

3.4. Loss Function

By simultaneously considering mask, class, and attribute results, we use bipartite matching as the cost. We use the focal loss [9] in mask prediction and classification. For mask prediction, we use dice loss [11]. As is the case with multi-label classification tasks, our model's predictions are supervised in a one-hot format. For attributes, we follow the default design of Attribute-Mask R-CNN [8]. It is an attribute loss with many binary labels. Here is how the entire loss might be expressed:

$$L = \sum_{j=1}^{3} \lambda_{cls}L_{cls}^{(j)} + \lambda_{mask}L_{mask}^{(j)} + \lambda_{attribute}L_{attribute}^{(j)} \tag{3}$$

where j means the training stage index and here all λ are empirically set to 1.

4. Experiments

4.1. Dataset, Metrics and Implementation Details

We conduct experiments on the Fashionpedia [8] dataset. The Fashionpedia dataset includes segmentation masks and

Method	Backbone	Schedule	Flops(B)	Params(M)	AP^{mask}_{IoU}	$AP^{mask}_{IoU+F_1}$
Attribute-Mask R-CNN [8]	R50-FPN	3x	296.7	46.4	39.2	29.5
	R101-FPN		374.3	65.4	40.7	31.4
	Swin-B		508.3	107.3	47.5	40.6
Fashionformer [13]	R50-FPN	3x	198.0	37.7	42.5	39.4
	R101-FPN		275.7	56.6	45.6	42.8
	Swin-B		442.5	100.6	49.5	46.5
Ours	R50-FPN	3x	196.4	36.5	45.4	39.8
	R101-FPN		270.2	53.5	48.2	43.4
	Swin-B		423.8	94.7	52.3	46.9

Table 1. Results on Fashionpedia.

attribute labels together, so it is appropriate for our model's training and inference.

For clothing instance segmentation, we adopt the mean average precision mAP^{mask}_{IoU} from the COCO setting. In order to evaluate clothing segmentation and attribute recognition together, we adopt mAP^{mask}_{IoU+F1} from [13], which is a joint metric combining mAP and $F1_{score}$. The GFlops are obtained with $3 \times 1020 \times 1020$ inputs following [8].

We implement our method using the PyTorch framework with MMDetection toolbox [2], and the model is trained on 2 NVIDIA RTX 3090 GPUs with the batch size of 16. One standard training schedule 1x is 5625 iterations, and we set the training epoch number as 3x. The learning rate is set by [7]. and both the generator and the discriminator are alternately updated in every iteration. We use large scale jittering [6] that resizes an image to a random ratio between [0.5, 2.0] of the target input image size.

4.2. Qualitative Evaluation

Fig. 3 compares visual results generated by our method and other SOTA models on R101-FPN backbone. Almost every attribute of the input image as well as flawless instance segmentation masks can be predicted by our model. Our model achieves the best segmentation results, especially for fashion items with large scale differences. Moreover, our method has better clothing segmentation and attribute results than other SOTA methods. As can be seen in Fig. 4, some tiny scale fashion items can be completely detected and segmented, and some small parts belonging to one clothing instance but scattered can also be detected and segmented into the correct instance. Furthermore, the attributes of the recognized instances are mostly complete and correct. It proves the effectiveness of the proposed ***multi-layer attention module*** in our decoder design.

4.3. Quantitative Evaluation

Table 1 compares our method with Fashionformer [13] and Attribute-Mask R-CNN [8] in different settings. By using R50-FPN backbone, our method outperforms Fashionformer and Attribute-Mask R-CNN, respectively, in terms of mAP^{mask}_{IoU} by 2.9% and 6.2%; and in terms of mAP^{mask}_{IoU+F1} by 0.4% and 10.3%. By using R101-FPN backbone, our method outperforms in terms of mAP^{mask}_{IoU} by 2.6% and 7.5%, and in terms of mAP^{mask}_{IoU+F1} by 0.6% and 12.0%. To compare with SpineNet [4] with more training iterations, Xu *et al.* [13] adopted Swin-B as the backbone and retrained their model. Comparatively, we also retrained on Swin-B backbone, our method still outperforms Fashionformer by 2.8% in mAP^{mask}_{IoU} and 0.4% in mAP^{mask}_{IoU+F1}, and is better than Attribute-Mask R-CNN by 4.8% in mAP^{mask}_{IoU} and 6.3% in mAP^{mask}_{IoU+F1}. This indicates the effectiveness of our proposed new ***multi-layer attention module*** in our decoder by aggregating features of fashion items with different scales.

5. Conclusion

In this paper, we design a new DETR-based model for the task of layered clothing segmentation and fine-grained attribute recognition. By refining the decoder with a specially designed ***multi-layered attention module***, we introduce an effective solution to the task, especially for cases when fashion items are of different scales. Extensive experiments on the Fashionpedia dataset have demonstrated that our method achieves competitive performance comparing with other state-of-the-art models.

Acknowledgements

The work is supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Grant Number 152112/19E) and by the Innovation and Technology Commission of Hong Kong, under grant ITP/028/21TP.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. 1, 2

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 4

[3] Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati, and Jiaying Liu. Fashion meets computer vision: A survey. *ACM Comput. Surv.*, 54(4), jul 2021. 1

[4] Xianzhi Du, Tsung-Yi Lin, Pengchong Jin, Golnaz Ghiasi, Mingxing Tan, Yin Cui, Quoc V Le, and Xiaodan Song. Spinenet: Learning scale-permuted backbone for recognition and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11592–11601, 2020. 4

[5] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5337–5345, 2019. 1, 2

[6] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 4

[7] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 4

[8] Menglin Jia, Mengyun Shi, Mikhail Sirotenko, Yin Cui, Claire Cardie, Bharath Hariharan, Hartwig Adam, and Serge Belongie. Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 316–332, Cham, 2020. Springer International Publishing. 1, 2, 3, 4

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[10] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2

[11] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 3

[12] Bowen Shi, Dongsheng Jiang, Xiaopeng Zhang, Han Li, Wenrui Dai, Junni Zou, Hongkai Xiong, and Qi Tian. A transformer-based decoder for semantic segmentation with multi-level context mining. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 624–639. Springer, 2022. 2

[13] Shilin Xu, Xiangtai Li, Jingbo Wang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. Fashionformer: A simple, effective and unified baseline for human fashion segmentation and recognition. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 545–563, Cham, 2022. Springer Nature Switzerland. 1, 2, 3, 4

[14] Shuai Zheng, Fan Yang, M. Hadi Kiapour, and Robinson Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 1670–1678, New York, NY, USA, 2018. Association for Computing Machinery. 1, 2