Fashionformer: A Simple, Effective and Unified Baseline for Human Fashion Segmentation and Recognition-Appendix

Shilin Xu[1,3*] Xiangtai Li[1,3] Jingbo Wang[2]
Guangliang Cheng[3] Yunhai Tong[1] Dacheng Tao[4]

[1] Key Laboratory of Machine Perception, MOE, School of Artificial Intelligence, Peking University
[2] CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[3] SenseTime Research
[4] The University of Sydney
lxtpku@pku.edu.cn, xushilin@stu.pku.edu.cn,chengguangliang@sensetime.com

Overview. In this supplementary, we provide the following information: more experimental details in experiment sections, more discussion on our work contribution and limitation, more visualization results on Fashionpedia, DeepFashion and ModaNet. We will open source all the model and code.

1 More Experimental Details and Discussion.

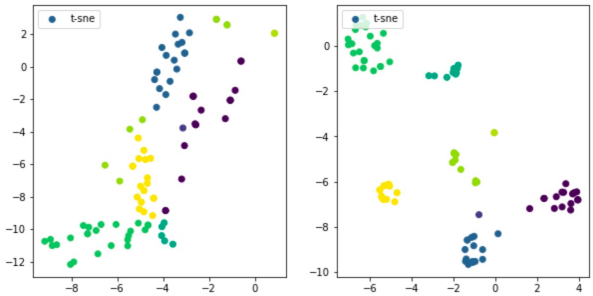
Detailed Training on Fashionpedia. Since this dataset is smaller than COCO, following original work, we enlarge the dataset by repeating several times of origin size to match the size of COCO dataset. We report standard 1× and 3× training schedules for fair comparison, following [4] where 1× is nearly 36 epochs. We use the same standard COCO data augmentation settings in mmdetection [1].

Training Settings on DeepFashion and ModaNet. We follow the main standard setting for training on both datasets. The main difference is we use the AdamW [5] with the initial learning rate 1e-4. We use multiscale training [1] where the short size is ranging from 800 to 1333. Since there is no clear benchmark for recently introduced models including CondInst [6], MaskFormer [2] and K-Net [7]. We re-implement all these methods in the same codebase. All the model use the single scale inference on both datasets.

Ablation on interaction number. We also increase the interaction number of our Fashionformer to 4 and 5. We find a minor performance gain (0.3%-0.5% $\text{mAP}^{\text{mask}}_{\text{IoU}}$, 0.2%-0.3% $\text{mAP}^{\text{mask}}_{\text{F1}}$) for 12 epochs training. However, when adopting more epochs training, there is no difference with interaction number 3. Thus, we set the default interaction number to 3.

Effectiveness of learned gating in Dynamic Convolution. Removing the gating in DC results in 1.0% $\text{mAP}^{\text{mask}}_{\text{IoU}}$ drop. This indicates the learned query should depend on the query feature conditionally rather simple addition.

Discussion on main contribution. *Our main contribution is to provide a new DETR-like baseline for fashion analysis including both fashion segmentation and*



(a) solo object queries (b) joint object queries

Fig. 1: Visual Comparison using T-SNE. The same color represents the same class. (We choose 10 classes).

its attribute prediction, which is a system-level invention. **One object query corresponds to one attribute query.** (Two embeddings.) The former is used to decode fashion masks, while the latter is used to predict the fashion attributes. Both queries decouple the two tasks which achieve better results than the shared query (see Tab1.(c) of the main paper).

To fulfil this goal, we present a two-stream query learning framework with in-dependent two queries. e denote the task is a complex and challenging, which contains instance segmentation and attribute prediction jointly. Compared with attribute Mask-RCNN which contains more heavily engineered heads including Box head, RPN head, Mask head and Attribute head, our method contains one two-stream head with two queries as outputs (without RPN, box head and attribute head). Thus, the pipeline is much simpler.

Benefits of Joint Training. In Tab.1(a) of the main paper, we show that adding attribute prediction leads to 3.1% mAP improvements compared with single object query baseline. We argue that fine attribute queries lead to better instance segmentation AP^{mask}_{IoU} **mainly on the classification** because attributes reduce the scope of classification (Line-083). We provide more evidences. We find the gaps on two models in cases of class labels and then randomly choose the 10 classes from top-20 labels. Then we perform T-SNE visualization on object queries for such classes (solo object queries on joint object and attribute queries). As shown in Fig. 1 (b), we show ***joint learning leads to more discriminative representation***.

Border Impact. Our work pushes the boundary of human fashion segmentation algorithms with simplicity and effectiveness. Compared with existing works, our unified baseline boosts the state-of-the-art models via a very large margin. Also, we believe that our model has huge potential for fashion retrieval, where the shopping companies may be interested.

Limitation and discussion. One limitation is that our method only considers a single scale feature X_{fuse} for the mask prediction. This leads to inferior results on small object predictions shown in the Tab.5 of the main paper. When compared

with SpineNet backbone [3]. However, our goal is to provide a new simple baseline rather than heavy engineering for better performance. One potential way to improve our framework is to consider multiscale mask prediction [8].

2 More visualization results.

More results on Fashionpedia. In Fig. 2, we show more visualization results on Fashionpedia dataset. We find our Fashionformer can obtain more consistent instance segmentation results with clear boundary.
More results on ModaNet. We show more comparison results on ModaNet in Fig. 3. Compared with recent works of instance level segmentation, our Fashionformer has clearly better results for both mask classification and high quality.

References

1. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
2. Cheng, B., Schwing, A.G., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. arXiv (2021)
3. Du, X., Lin, T.Y., Jin, P., Ghiasi, G., Tan, M., Cui, Y., Le, Q.V., Song, X.: Spinenet: Learning scale-permuted backbone for recognition and localization. In: CVPR. pp. 11592–11601 (2020)
4. Jia, M., Shi, M., Sirotenko, M., Cui, Y., Cardie, C., Hariharan, B., Adam, H., Belongie, S.: Fashionpedia: Ontology, segmentation, and an attribute localization dataset. In: ECCV. pp. 316–332. Springer (2020)
5. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2017)
6. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. arXiv preprint arXiv:2003.05664 (2020)
7. Zhang, W., Pang, J., Chen, K., Loy, C.C.: K-net: Towards unified image segmentation. NeurIPS (2021)
8. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. ICLR (2020)

Fig. 2: More Visual comparison results on Fashionpedia using ResNet-50 backbone. Best view in color.

Fig. 3: More Visual comparison results on ModaNet using ResNet-50 backbone. Best view in color.