Dress like an Internet Celebrity: Fashion Retrieval in Videos

Hongrui Zhao[1] , **Jin Yu**[2] , **Yanan Li**[3] , **Donghui Wang**[1*] , **Jie Liu**[2] , **Hongxia Yang**[2] and **Fei Wu**[1]

[1]College of Computer Science and Technology, Zhejiang University

[2]Alibaba Group

[3]Institute of Artificial Intelligence, Zhejiang Lab

{hrzhao, ynli, dhwang, wufei}@zju.edu.cn, {kola.yu, sanshuai.lj, yang.yhx}.alibaba-inc.com

Abstract

Nowadays, both online shopping and video sharing have grown exponentially. Although internet celebrities in videos are ideal exhibition for fashion corporations to sell their products, audiences do not always know where to buy fashion products in videos, which is a cross-domain problem called video-to-shop. In this paper, we propose a novel deep neural network, called Detect, Pick, and Retrieval Network (DPRNet), to break the gap between fashion products from videos and audiences. For the video side, we have modified the traditional object detector, which automatically picks out the best object proposals for every commodity in videos without duplication, to promote the performance of the video-to-shop task. For the fashion retrieval side, a simple but effective multi-task loss network obtains new state-of-the-art results on DeepFashion. Extensive experiments conducted on a new large-scale cross-domain video-to-shop dataset show that DPRNet is efficient and outperforms the state-of-the-art methods on video-to-shop task.

1 Introduction

With the rapid development of e-commerce, people are more and more used to shopping online. According to the statistics, retail e-commerce sales worldwide amounted to 3.53 trillion US dollars in 2019. Besides, the total gross merchandise volume (GMV) is 38.3 billion U.S. dollars for just 24 hours on Alibaba's double eleven shopping carnival in 2019. Meanwhile, video sharing application and live video streaming have become increasingly popular in recent years. When watching internet celebrities on Taobao Live, Tik Tok, and YouTube, etc, audiences may be interested in their fashion collocation and willing to buy what the internet celebrities wear. Videos show a great marketing ability for e-commerce, as they can influence thousands of people and stimulate people's desire to buy the same commodity in videos. To promote product sales, finding the clothing in videos to the same items in e-commerce websites is part-and-parcel of online shopping.

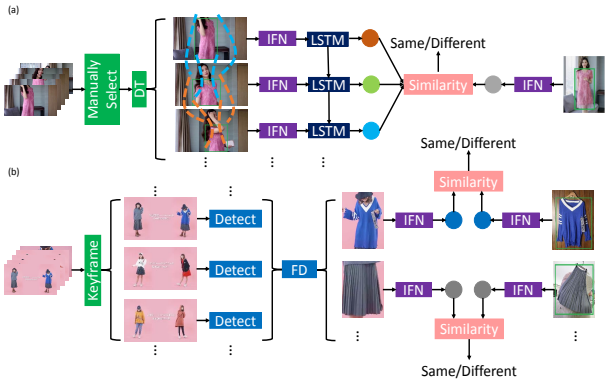

Figure 1: (a) shows the previous video-to-shop pipeline, which manually extracts frames containing the corresponding clothing in videos as the clothing trajectories. Then, (a) doing DT (detection and tracking) on clothing trajectories. Finally, (a) formulating retrieval as a multiple-to-single matching problem after feature construction using IFN (image feature network) and LSTM. (b) shows our pipeline, which automatically picks out an optimal visual quality proposal for each clothing instance from key frames in videos without duplication by FD (Filter and Deduplication), and formulates retrieval as a single-to-single matching problem. Less is more!

Although consumer-to-shop clothing retrieval has made great progress [Liu *et al.*, 2016b; Ge *et al.*, 2019; Kuang *et al.*, 2019], etc, which finds the same clothing in online shops by photos from consumers, searching the same clothing from videos in the online shops has few studies yet. Compared with consumer-to-shop, video-to-shop is more difficult due to various viewpoints, occlusion, and crop for captured clothing in videos. The above variations of the same clothing may cause significant visual discrepancy, and undoubtedly greatly degrade fashion retrieval performance [Ge *et al.*, 2019; Kuang *et al.*, 2019]. Another problem is how to pick out all the clothing proposals in the video, and accurately classify them into instance-level. AsymNet [Cheng *et al.*, 2017] solves the above two problems by manually extracting a constant number of frames, which contains the corresponding clothing, from videos as clothing trajectories. Then, AsymNet employs a clothing detector and object tracker to generate multiple clothing proposals. Finally, AsymNet formulates

*Corresponding Author.

fashion retrieval as a multiple-to-single matching problem using the feature from IFN and LSTM (see Fig. 1(a)). Artificial assistance limits its large-scale application. What's more, multiple-to-single matching may not only bring in noisy features caused by bad viewpoint, occlusion, and crop for the same clothing instance in videos but also increase time cost due to use IFN and LSTM multiple times. We argue that only an optimal visual quality (frontal viewpoint, without cropping and no or slight occlusion) clothing proposal is enough to retrieval the same item in the gallery precisely and fast.

Thus, we propose the Detect, Pick, and Retrieval Network (DPRNet, see Fig. 1(b)). On the video side, DPRNet employs a key frame mechanism and only does video clothing detection on key frames. Then, DPRNet filters the bad visual quality proposals and classifies the remaining proposals into instance-level according to the distance of their feature embeddings. Eventually, DPRNet outputs an optimal visual quality proposal for each clothing instance in the video. On the fashion retrieval side, we propose a multi-task loss image feature network which is simple but effective. Thanks to the above mechanisms, DPRNet only needs around two weeks to automatically accomplish the video-to-shop task on 10 million videos and hundreds of millions of gallery images with 200 GPUs.

The main contributions of this work are as follows:

- A novel deep-based network, DPRNet, which consists of three components: video clothing detection, clothing proposal picking, and fashion retrieval, is proposed for video-to-shop application. It offers an automatic solution for the large-scale online video-to-shop task.
- For video clothing detection and picking, the proposal scoring mechanism can select optimal quality clothing proposals and significantly improve the video-to-shop task performance. Meanwhile, a deduplication mechanism makes video-to-shop more efficient.
- Extensive experiments have been conducted on a new large-scale video-to-shop dataset, which consists of 818 videos and 21614 gallery images, to evaluate the performance of the proposed DPRNet. Experiments show that DPRNet achieves excellent performance and outperforms the existing state-of-the-art methods.

2 Related Work

2.1 Image and Video Object Detection

State-of-the-art image object detection methods mostly follow two paradigms, two-stage and one-stage. A two-stage pipeline (*e.g.* R-CNN [Girshick *et al.*, 2014], Fast R-CNN [Girshick, 2015], Faster R-CNN [Ren *et al.*, 2015], etc.) firstly generates region proposals, and then classifies and refines each proposed bounding box. Comparing to two-stage detectors, one-stage detectors directly predict the bounding box of interest based on the extracted feature map from CNN, which is more efficient but may hinder the performance a little bit. Representative works include YOLO [Redmon *et al.*, 2016] and its variants [Redmon and Farhadi, 2017; Redmon and Farhadi, 2018], SSD [Liu *et al.*, 2016a] and RetinaNet [Lin *et al.*, 2017]. However, it is difficult to extend one-stage detectors to more complicated tasks. Thus, we take Faster R-CNN as our basic detector to make use of proposal-level object semantic features in our following work.

Compared with image object detection, video object detection benefits less from deep learning until the emergence of the ImageNet VID dataset. DFF [Zhu *et al.*, 2017] proposed an efficient framework, which runs the expensive convolutional network only on sparse key frames. This method achieves 10x speedup than per-frame evaluation with moderate accuracy loss. [Chen *et al.*, 2018] proposed a Scale-Time Lattice, a flexible framework that offers a rich design space to balance the performance and cost in video object detection. Inspired by the above works, we also adopt a key frame mechanism to speed up video clothing detection.

2.2 Fashion Retrieval

With the development of convolutional neural networks, fashion retrieval has made great progress [Huang *et al.*, 2015; Hadi Kiapour *et al.*, 2015; Liu *et al.*, 2016b; Gajic and Baldrich, 2018; Zhang *et al.*, 2018; Dodds *et al.*, 2018; Chopra *et al.*, 2019; Ge *et al.*, 2019; Kuang *et al.*, 2019]. [Liu *et al.*, 2016b; Ge *et al.*, 2019] utilized multi-task learning strategy to learn feature representations. [Gajic and Baldrich, 2018] trained a Siamese network with standard triplet loss. [Dodds *et al.*, 2018] trained triplet-based network with an effective online sampling technique. [Chopra *et al.*, 2019] proposed a Grid Search Network for learning feature embeddings for fashion retrieval. [Kuang *et al.*, 2019] proposed a Graph Reasoning Network (GRNet), which first represents the multi-scale regional similarities and their relationships as a graph.

Despite fashion retrieval in consumer-to-shop has been pushed into a new phase, there are few studies focused on finding clothing in videos to the same items in online shops. AsymNet is proposed in [Cheng *et al.*, 2017], which attempted to exact match clothing in videos to online shops. Different from previous work, our DPRNet automatically picks out optimal visual quality clothing proposals without duplication in videos and formulates fashion retrieval a single-to-single matching problem.

3 Our Method

Fig. 2 illustrates the overview of DPRNet. In this section, we will elaborate on the details of DPRNet.

3.1 Video Clothing Detection

Proposal scoring is conceptually simple: Faster R-CNN with additional branches that output the visual quality scores of the clothing proposals, which will help DPRNet to pick out optimal visual quality clothing proposals in videos to improve the performance of the video-to-shop task. The following sections are the details of the proposal scoring branches.

Above all, let us briefly review Faster R-CNN [Ren *et al.*, 2015]. Faster R-CNN consists of two stages. The first stage is called Region Proposal Network (RPN), which proposes candidate object bounding boxes regardless of object categories. The second stage is extracting features using RoIAlign for each proposal and performing proposal classification and bounding box regression. Since Faster R-CNN is

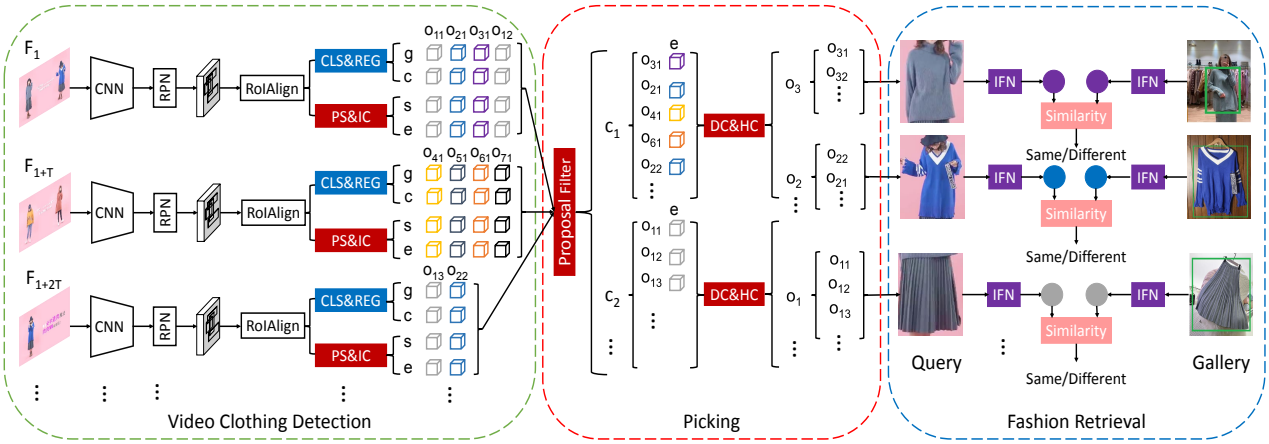

Figure 2: An overview of Detect, Pick, and Retrieval Networks (DPRNet) for video-to-shop clothing retrieval. Given a video, Region Proposal Network (RPN) is first employed to produce clothing proposals from sparsely selected key frames at a fixed interval of T. After that, the CLS (Classification) and REG (Regression) branches are devised to get bounding box g and class c of each proposal after RoIAlign ($\mathbf{o}_{ij}$: i denotes the i-th clothing instance and j denotes the j-th proposal of instance i). Meanwhile, PS (Proposal Scoring) and IC (Instance Classification) branches output the quality (viewpoint, occlusion, and crop) scores s of each clothing proposal and instance-level proposal feature embeddings e, respectively. To improve the performance of deduplication and clothing retrieval, DPRNet will filter low quality score proposals. Then, DC (Distance Calculating) and HC (Hierarchical Clustering) classify the proposals into different instances according to their feature embeddings output from IC branch. For proposals of each instance, DPRNet will retain the highest quality score proposal. Finally, the retained proposals serve as the query images and find the same item in the gallery.

a fully differentiable model, it can be trained end-to-end with the following loss function:

$$L_{fasterrcnn} = L_{rpn\_cls} + L_{rpn\_reg} + L_{rcnn\_cls} + L_{rcnn\_reg}, \quad (1)$$

where L_{rcnn_cls} and L_{rcnn_reg} are the classification loss and bounding box regression loss of R-CNN. Here, classification loss is the cross-entropy loss, and regression loss is the smooth-L1 loss. And, Region Proposal Network (RPN) has the same loss function as R-CNN.

The proposal scoring branches aim to evaluate the visual quality of clothing proposals. And, we formulate the proposal scoring as a specific supervised classification problem. We use the same two-stage strategy, with an identical Region Proposal Network (RPN) in the first stage. In the second stage, in parallel to output the class and box offset, proposal scoring branches also predict occlusion, viewpoint, and crop score for each RoI respectively, as shown in Fig. 3. The three proposal scoring branches consist of 4 convolution layers and 1 fully connected layers respectively. For the 4 convolution layers, we follow the RCNN head and set the kernel size and filter number to 3 and 1024 respectively for all the convolution layers. For the fully connected (fc) layer, we also follow the RCNN head and set the input of the fc layer to 1024 after average pooling layer and the output of the fc to $C_{occlusion}$, $C_{viewpoint}$ and C_{crop}, which represent the number of occlusion, viewpoint and crop classes respectively. Formally, during training, we define a proposal visual quality loss on each sampled RoI as:

$$L_{quality} = L_{occlusion} + L_{viewpoint} + L_{crop}, \quad (2)$$

because the proposal scoring have been formulated as a specific supervised learning problem of classification, we define

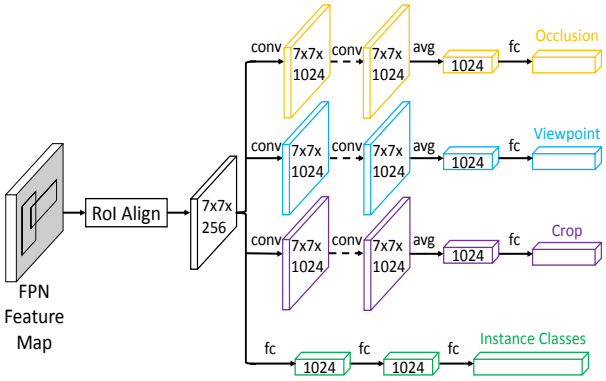

Figure 3: Proposal Scoring and Instance Classification Branches.

$L_{occlusion}$, $L_{viewpoint}$ and L_{crop} as cross-entropy loss.

To remove duplicate proposals, we add an instance classification branch, as shown in Fig. 3. The instance classification branch has 3 fully connected (fc) layers. The input of the first fc is 1024 and the output of the last fc is $C_{instance}$, which represents the total number of clothing instances in the training set. Thus, the $L_{instance}$ is a cross-entropy loss. During inference, the last but one fc output is used as feature embeddings to calculate the similarities of different clothing proposals.

Finally, for the proposed detector with Proposal Scoring and Instance Classification, the overall loss formulation is:

$$L_{detection} = L_{fasterrcnn} + L_{quality} + L_{instance}, \quad (3)$$

where $L_{fasterrcnn}$ is the loss of Faster R-CNN in Eq.(1), $L_{quality}$ is the loss for the proposed proposal scoring branches and $L_{instance}$ is the instance classification loss.

3.2 Picking (Proposal Filter and Deduplication)

After video clothing detection, picking out optimal visual quality proposals and removing duplicate proposals are critical to improving the performance of video-to-shop fashion retrieval. Firstly, we filter bad visual quality proposals according to their proposal scores. If a proposal satisfies $s_{crop} > \theta_{crop}$, $s_{occlusion} > \theta_{occlusion}$ and $s_{viewpoint} > \theta_{viewpoint}$ in the same time, we will retain the proposal, where θ_{crop}, $\theta_{occlusion}$ and $\theta_{viewpoint}$ are the thresholds. Then, we will send the retained proposals to the proposal pool according to their coarse-grained class. And, we use the instance classification branch feature embeddings e to calculate proposal dissimilarity in each coarse-grained class proposal candidate pool by cosine similarity as follow:

$$dis(e_i, e_j) = 1 - \frac{e_i \cdot e_j}{\|e_i\| \, \|e_j\|}. \tag{4}$$

After that, based on the hierarchical clustering theory [Müllner, 2011], we utilize the agglomerative hierarchical clustering strategy to classify the proposals into instance-level by their dissimilarity distance. For the same instance proposals, we find the highest quality score proposal as our output and do fashion retrieval.

3.3 Fashion Retrieval Network (IFN)

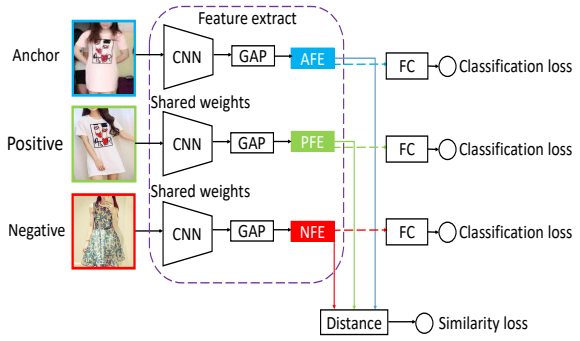

Figure 4: The framework of the proposed multi-task learning deep network for fashion retrieval. GAP denotes Global AvgPooling and FC denotes the Fully Connected layer. AFE, PFE, and NFE denote anchor feature embeddings, positive feature embeddings, and negative feature embeddings, respectively.

To deal with the fashion retrieval task, we propose a simple but effective fashion retrieval network with multi-task learning techniques. The proposed network, as shown in Fig. 4, has a CNN backbone network and two branches. The distance branch is to measure the similarities of clothing instances and make the feature embeddings of the same instances closer while the feature embeddings of the different instances farther. The classification branch is to distinguish different categories and encourage the learned features to be discriminative. The distance branch is a triplet-based model. The standard triplet loss is proposed in FaceNet [Schroff *et al.*, 2015]:

$$\begin{aligned} L_{triplet} = \sum_{i}^{N} [& \|f\left(x_i^a\right) - f\left(x_i^p\right)\|_2^2 \\ & - \|f\left(x_i^a\right) - f\left(x_i^n\right)\|_2^2 + \alpha]_+, \end{aligned} \tag{5}$$

where α is a margin that is enforced between positive and negative pairs, N is the number of triplets. x is the input image and $f(x)$ is the feature representation of image x. However, the disadvantage of standard triplet loss is that triplets are randomly selected from the training set, so it may be a simple combination of samples, that is, very similar positive samples and very different negative samples. Thus, we adopt the "batch-hard" sampling technique proposed in [Hermans *et al.*, 2017], which concentrates on the person re-ID problem. Thus, the loss is defined as:

$$\begin{aligned} L_{batchhard} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{all\ anchors} [& \overbrace{\max_{p=1\ldots K} \|f(x_i^a) - f(x_i^p)\|_2^2}^{hardest\ postive} \\ & - \underbrace{\min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j\neq i}} \|f(x_i^a) - f(x_i^n)\|_2^2}_{hardest\ negtive} + \alpha]_+, \end{aligned} \tag{6}$$

P is the number of sampled classes (number of clothing instances), and K is the number of images randomly sampled for each class (instance), thus making up a batch of PK images. For the classification branch, we use a label smooth softmax loss in [Szegedy *et al.*, 2016]:

$$L_{cls\_ls} = \sum_{i}^{N} -(q_i) log(p_i) \begin{cases} 1 - \frac{N-1}{N}\varepsilon, y = i \\ \frac{\varepsilon}{N}, y \neq i \end{cases}, \tag{7}$$

where N is the number of instances. Given an image, we denote y as ground truth label and p_i as label prediction logits of instance i. ε is a small constant to encourage the model to be less confident on the train set. Thus, our network total losses are as follow:

$$L_{retrieval} = L_{cls\_ls} + L_{batchhard} \tag{8}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

DeepFashion. Deepfashion [Liu *et al.*, 2016b] provides over 800,000 real-world images with rich additional information about categories, attributes, landmarks, etc. Besides, DeepFashion consumer-to-shop retrieval is a popular dataset for evaluating cross-domain fashion retrieval. Thus, we train and evaluate the fashion retrieval network on DeepFashion consumer-to-shop retrieval dataset.

DeepFashion2. Due to the DeepFashion was limited by single clothing-item per image, we use the training set of DeepFashion2 [Ge *et al.*, 2019], which contains 191,961 images annotated with a bounding box and category label for each clothing item, to train the clothing detector. Besides, the clothing items in the training set are also annotated with viewpoint (no wear, frontal, and side or back) and occlusion (slight, medium, and heavy). To solve the problem of lacking crop label, we annotate the integrity of each clothing item based on whether their key landmarks are labeled. Thus, we get all the labels to train the proposal scoring branches. Finally, we also generate an instance-level class label for each clothing instance to train the instance classification branch.

Figure 5: Proposal scoring results from DPRNet. Each row of images is from the same video.

Video-to-Shop Dataset. To evaluate the performance of DPRNet, we collect a new large-scale dataset for the video-to-shop task. Taobao.com and Tmall.com have a lot of online shops sell the same clothing items that appear in Taobao Live, Tik Tok, etc. Besides, these online shops show videos and images for the same clothing items simultaneously. We download these video and image pairs from Alibaba Group. To expand the number of gallery images and make video-to-shop fashion retrieval more difficult, we add part of images from DeepFashion to the gallery. Comparison of the video-to-shop datasets is shown in Tab. 1.

For video clothing detection, we need to evaluate the picking, filter, and deduplication ability of DPRNet. We will calculate R_{pick}, $R_{duplication}$ and $R_{optimal}$ results on the video-to-shop dataset manually to evaluate the performance of video clothing detection. The mathematical formula is as follows:

$$R_{pick} = \frac{\sum_{i \in I} \mathbb{I}\{pick(i)\}}{I}, \tag{9}$$

$$R_{duplication} = \frac{\sum_{j \in I_p} \mathbb{I}\{duplication(j)\}}{I_p}, \tag{10}$$

$$R_{optimal} = \frac{\sum_{j \in I_p} \mathbb{I}\{optimal(j)\}}{I_p}, \tag{11}$$

where I and $I_p = \sum_{i \in I} \mathbb{I}\{pick(i)\}$ is the total number of clothing instances and picked out clothing instances, respectively. $pick(i)$ denotes whether picking out the i-th clothing instance in videos. $duplication(j)$ denotes whether repeatedly picking out the j-th instance in the set of picked out clothing instances. $optimal(j)$ denotes whether the proposal of the j-th instance is optimal visual quality in the set of picked out clothing instances.

For the fashion retrieval evaluation, because it is a typical information retrieval task, we use Recall@K metric to measure the performance.

4.2 Performance of Video Clothing Detection and Picking

We conduct experiments on our video-to-shop dataset to study the performance of video clothing detection by the metrics defined above when setting different key frame intervals. The results are shown in Fig. 6. We can see that DPRNet gets excellent performance to automatically pick out clothing proposals in videos without redundant. What's more, even with a large key frame interval, DPRNet only endures a little performance loss. Therefore, DPRNet is effective and efficient enough to be applied in real scenarios.

Datasets	AsymNet	Ours
Videos	526	818
Query	5266	5013
Gallery	17128	21614

Table 1: Comparison of the video-to-shop retrieval datasets.

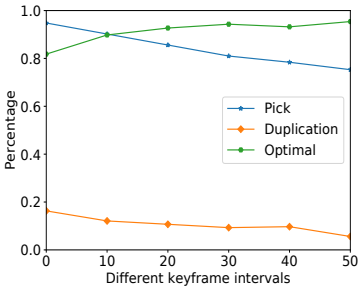

Figure 6: The performance of DPRNet when setting different key frame intervals.

In Fig. 5, we show some examples of proposal scoring results from DPRNet. We can see that proposal scoring branches can give an accurate quality score of clothing proposals. This illustrates the proposal scoring branches can help us to find out optimal visual quality clothing proposals in videos. Undoubtedly, optimal visual quality clothing proposals will promote the performance of fashion retrieval [Kuang *et al.*, 2019; Ge *et al.*, 2019].

4.3 Performance of Fashion Retrieval Networks

Tab. 2 shows the detailed performance of our method with state-of-the-art methods. Our method outperforms previous state-of-the-art methods in both val+test (95,961 query images and 22,669 gallery images) and test (47,434 query images and 11,312 gallery images) split ways. Notably, GRNet introduces the similarity pyramid with graph reasoning. Our method is only based on a simple CNN backbone without bells and whistles.

In Fig. 7, we show some results of the fashion retrieval network on the DeepFashion consumer-to-shop benchmark.

Methods	Top-1	Top-20	Top-50
FashionNet [Liu *et al.*, 2016b]	7.0	18.8	22.8
VAM+ImgDrop [Wang *et al.*, 2017]	13.7	43.9	56.9
Triplet-SS [Gajic and Baldrich, 2018]	16.0	45.0	54.0
DREML [Xuan *et al.*, 2018; Kuang *et al.*, 2019]	18.6	51.0	59.1
KPM [Shen *et al.*, 2018; Kuang *et al.*, 2019]	21.3	54.1	65.2
NegTriplet [Dodds *et al.*, 2018]	23.0	60.9	72.0
GSN [Chopra *et al.*, 2019]	25.0	47.0	57.0
GRNet [Kuang *et al.*, 2019]	25.7	64.4	75.0
Ours(Val+Test)	**26.8**	**66.0**	**75.5**
Ours(Test)	**34.2**	**72.3**	**80.7**

Table 2: Comparison with state-of-the-art methods on DeepFashion consumer-to-shop benchmark.

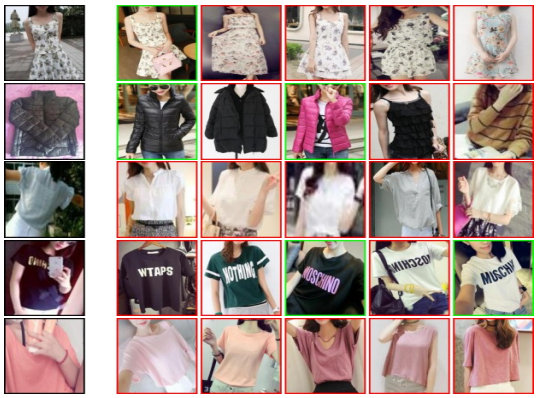

Figure 7: Example with top-5 retrieval results on Deepfashion consumer-to-shop benchmark.

Our retrieval model has an excellent ability to understand the clothing instance. However, the failure cases demonstrate that occlusion (4th row), bad viewpoint (3rd row), and incomplete (5th row) query clothing images degrade the performance of fashion retrieval. Fortunately, in the video-to-shop task, proposal scoring branches can help DPRNet to find out optimal visual quality clothing proposals, which would promote the performance of the video-to-shop task greatly.

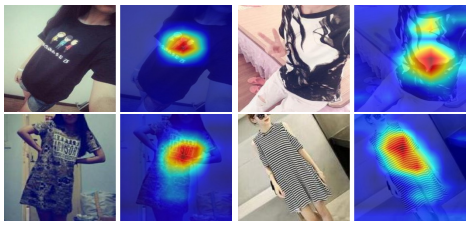

Figure 8: Visualization of important regions in images.

In Fig. 8, we utilize the attention map to visualize the important regions when doing inference. This illustrates that the fashion retrieval network will automatically focus on the discriminative area in images, such as pattern, text, and texture, etc.

4.4 Evaluation of DPRNet

To verify the effectiveness of the proposed single-to-single retrieval method of DPRNet, we compare it with the following multiple-to-single video-to-shop fashion retrieval methods: average (Avg), maximum (Max), Last and AsymNet [Cheng *et al.*, 2017][1]. A comparison of retrieval performance is shown in Fig. 9. We can see that the performance of AsymNet is better than Avg, Max, and Last. The main reason is AsymNet formulates multiple-to-single as a mixture estimation problem and gives different queries a weight. However, AsymNet still can't overcome the aggregate noise introduced by occlusion, crop, and side or back viewpoint clothing proposals. Thus, DPRNet outperforms the above methods with an impressive margin. What's more, ablation experiments investigate the effectiveness of different scoring branches in the proposed DPRNet is showed in Fig. 10.

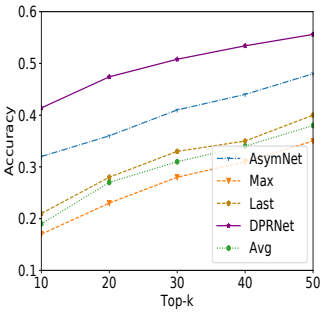

Figure 9: Comparison of video-to-shop retrieval methods.

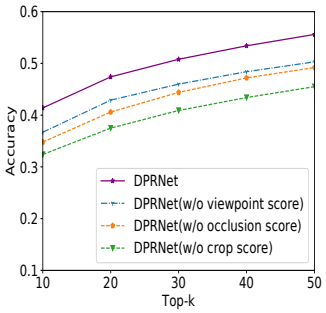

Figure 10: Ablation study of scoring branch.

In Tab. 3, we compare the time cost of feature construction for query instances from videos and gallery images. We can see without multiple-to-single and LSTM strategy, our method is faster than AsymNet. As for video clothing proposal picking, our method gives an automatic solution instead of human assistance in AsymNet.

	query	gallery
AsymNet	0.625 instance/sec	200 images/sec
Ours	850 instance/sec	850 images/sec

Table 3: Comparison of speed for image feature construction.

5 Conclusion

In this paper, we focus on the practical problem of video-to-shop application. Our method not only improves the performance of fashion retrieval but also makes the video-to-shop application faster and automatically. Especially, we wish that the idea of finding an optimal visual quality proposal in the video may shed light on other video-to-image tasks other than the video-to-shop application.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.61473256), the Natural Science Foundation of Zhejiang Province (No.LQ20F030007), the Youth Science Foundation of Zhejiang Lab (No.2020KD0AA03), and a research fund supported by Alibaba Group.

[1]Results from the different dataset are compared in Tab. 1.

References

[Chen *et al.*, 2018] Kai Chen, Jiaqi Wang, Shuo Yang, Xingcheng Zhang, Yuanjun Xiong, Chen Change Loy, and Dahua Lin. Optimizing video object detection via a scale-time lattice. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7814–7823, June 2018.

[Cheng *et al.*, 2017] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. Video2shop: Exact matching clothes in videos to online shopping images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4048–4056, July 2017.

[Chopra *et al.*, 2019] Ayush Chopra, Abhishek Sinha, Hiresh Gupta, Mausoom Sarkar, Kumar Ayush, and Balaji Krishnamurthy. Powering robust fashion retrieval with information rich feature embeddings. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[Dodds *et al.*, 2018] Eric Dodds, Huy Nguyen, Simao Herdade, Jack Culpepper, Andrew Kae, and Pierre Garrigues. Learning embeddings for product visual search with triplet loss and online sampling. *CoRR*, abs/1810.04652, 2018.

[Gajic and Baldrich, 2018] Bojana Gajic and Ramon Baldrich. Cross-domain fashion image retrieval. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1869–1871, June 2018.

[Ge *et al.*, 2019] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5337–5345, June 2019.

[Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, June 2014.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, December 2015.

[Hadi Kiapour *et al.*, 2015] M. Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, and Tamara L. Berg. Where to buy it: Matching street clothing photos in online shops. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 3343–3351, December 2015.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

[Huang *et al.*, 2015] Junshi Huang, Rogerio S. Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1062–1070, December 2015.

[Kuang *et al.*, 2019] Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 3066–3075, October 2019.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Oct 2017.

[Liu *et al.*, 2016a] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37, 2016.

[Liu *et al.*, 2016b] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, June 2016.

[Müllner, 2011] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms, 2011.

[Redmon and Farhadi, 2017] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7263–7271, July 2017.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[Redmon *et al.*, 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, June 2016.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*, pages 91–99. 2015.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015.

[Shen *et al.*, 2018] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-end deep kronecker-product matching for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6886–6895, June 2018.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, June 2016.

[Wang *et al.*, 2017] Z. Wang, Y. Gu, Y. Zhang, J. Zhou, and X. Gu. Clothing retrieval with visual attention model. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2017.

[Xuan *et al.*, 2018] Hong Xuan, Richard Souvenir, and Robert Pless. Deep randomized ensembles for metric learning. In *The European Conference on Computer Vision (ECCV)*, pages 723–734, September 2018.

[Zhang *et al.*, 2018] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. Visual search at alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery; Data Mining*, KDD '18, pages 993–1001, 2018.

[Zhu *et al.*, 2017] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2349–2358, July 2017.