DeepMark: One-Shot Clothing Detection

Alexey Sidnev[1,2], Alexey Trushkov[1], Maxim Kazakov[1,3], Ivan Korolev[1], Vladislav Sorokin[1]

[1]Huawei Research Center, Nizhny Novgorod, Russia
[2]Lobachevsky State University of Nizhny Novgorod, Russia
[3]National Research University Higher School of Economics, Nizhny Novgorod, Russia
{*sidnev.alexey, trushkov.alexey, kazakov.maxim, korolev.ivan, sorokin.vladislav*}*@huawei.com*

Abstract

The one-shot approach, DeepMark, for fast clothing detection as a modification of a multi-target network, CenterNet [1], is proposed in the paper. The state-of-the-art accuracy of 0.723 mAP for bounding box detection task and 0.532 mAP for landmark detection task on the DeepFashion2 Challenge dataset [2] were achieved. The proposed architecture can be used effectively on the low-power devices.

1. Introduction

Fashion image analysis and processing are highly demanded tasks and there are some popular services that require it [3], [4], [5], [6]. For example, virtual fitting rooms are becoming very common [7], [8].

Fashion analysis is quite a complex task. It requires solving of the following problems: detection, fine-gained classification (attributes), image retrieval, and overlay (augmented reality). Clothing detection is the first technological challenge to be solved for almost any fashion analysis task. Identifying the location of clothing items can be done with the different levels of precision: bounding boxes, keypoints, or segmentation. The first two tasks can provide enough information for the most of the real-life use-cases.

Data safety policies and customer needs arise another challenge requiring all processing to be done on the low-power edge devices like mobile phones and smart cameras. In accordance with the proposed approach, the tasks of bounding box prediction and keypoint estimation can be performed with a single one-shot neural network. It can also be used effectively on low-power devices without any post-processing work needed.
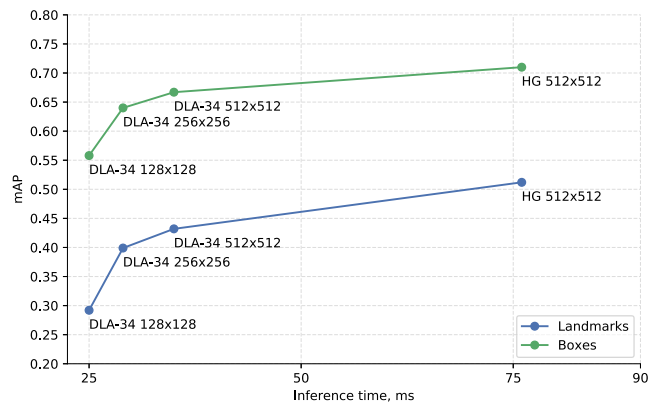

Figure 1. Speed-accuracy trade-off for object detection and landmark estimation on DeepFashion2 Challenge dataset [2]. We considered different image resolutions: 128×128, 256×256 and 512×512. DLA-34 model was initially trained on 512×512 images and then fine-tuned for other resolutions. Inference time was evaluated without flip and multiscale post processing.

2. Related work

Visual analysis of fashion images includes a variety of tasks: fashion attribute prediction; clothing recognition, detection and segmentation; estimation of fashion landmarks, clothing image retrieval and identification of fashion trends [9], [10], [11], [12], [13]. Previously, the different ways of distinctive features capturing in the fashion domain have been researched: full image usage [14], general object proposal [14], bounding boxes [15], masks [16], fashion attributes [13] and fashion landmarks [17].

Recent research shows that fashion landmarks are the most efficient, discriminative and robust representation for visual fashion analysis [13], [17].

Recent CNN-based approaches [13], [17], [18], [16], [19] along with the advent of large-scale fashion datasets [13], [11] have significantly outperformed the prior work.

The success of those deep learning based fashion models has demonstrated the strong representative power of neural networks and the advantages of fashion landmarks as features for fashion analysis [17].

However, until the release of the newest DeepFashion2 [20] dataset the full richness of domain-specific clothing knowledge had remained unresearched. Previously, the largest existing fashion dataset DeepFashion [13] was limited by a single clothing item per image, few landmarks and pose definitions.

DeepFashion2 [20] addressed such drawbacks providing a large-scale benchmark with comprehensive tasks and annotations. DeepFashion2 enabled the development of more advanced algorithms for visual fashion analysis.

In the paper, the tasks of clothing landmark prediction and clothing detection using DeepFashion2 Challenge [2] dataset are focused on.

The previous methods in the area in question describe detection of joints and other landmarks for facial alignment [21] and estimation of human body poses [22]. There are also some new works which followed the keypoint estimation approach for object detection. CornerNet [23] detects two bounding box corners as keypoints while ExtremeNet [24] detects the top-, left-, bottom-, rightmost; and center points of all objects. There is also CenterNet, which extracts a single center point per object [1].

3. Proposed approach

Due to the lack of existing works in fashion landmark detection, approaches from the similar fields can be good research-subjects (e.g. multi-person pose estimation [1], [25], [26], [27], [28]). Recently, CenterNet [1] has achieved the best speed-accuracy trade-off in object detection and pose estimation tasks on the MS COCO dataset. It verifies that CenterNet is a general approach that can be adapted to a new task, namely to fashion landmark detection. The primary challenge resides in fact, that CenterNet has been trained and has demonstrated solid results on the MS COCO pose estimation dataset, which includes only one class of objects (human) and a fixed number of keypoints (17). The DeepFashion2 [20] dataset includes 13 categories each of which contains a different number of keypoints (294 keypoints total). Our approach is based on CenterNet topology but handles two above mentioned restrictions with ease.

Fig.2 depicts the scheme of our DeepMark approach. The DeepFashion2 dataset contains 13 classes, so 13 channels are used to predict the centers of the boxes. Width and height are predicted directly. Two more channels in the output feature map representing $\triangle x$ and $\triangle y$ are regressed to refine the center coordinates.

Fashion landmark estimation aims to estimate 2D keypoint locations for each item of clothing in one image. To predict the keypoints, regression-based estimator similar to
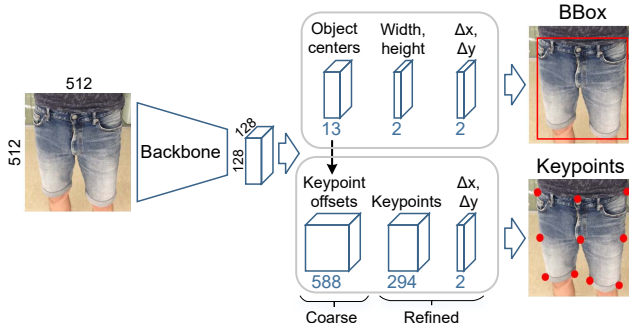

Figure 2. Scheme of DeepMark.

Subset	Challenge	Paper
Train	191961	390884
Val	32153	33669

Table 1. Training and validation subsets provided for DeepFashion2 Challenge [2] and used in original paper [20].

Metric	DeepMark	Challenge	Paper
mAP_{pt}	0.601	**0.605**	0.641
	0.532	0.529	0.563
$mAP_{pt}^{iou=0.50}$	**0.793**	0.790	0.820
	0.784	0.775	0.805
$mAP_{pt}^{iou=0.75}$	0.683	**0.684**	0.728
	0.599	0.596	0.641

Table 2. Landmark estimation results for DeepMark and Mask R-CNN[20]. The proposed approach outperforms DeepFashion2 Challenge[2] baseline on the almost all metrics. Results of evaluation on visible landmarks only and evaluation on both visible and occluded landmarks are separately shown in each row.

[29] is used. The centers of the boxes are used to predict the coarse locations of the keypoints (keypoint offsets from Fig.2). To refine the estimation of keypoints, a heatmap is generated for each of 294 keypoints. Like in the detection case, two more channels $\triangle x$ and $\triangle y$ are used to get more precise landmark coordinates. In order to preserve connectivity between object center and corresponding keypoints, each refined keypoint position is assigned to its closest coarse position from the previous step. These refined coordinates represent the final results of the algorithm.

4. Results

All experiments were performed on the publicly available DeepFashion2 Challenge dataset [2], which contains 191,961 images in the training set and 32,153 images in the validation set. The previous work [20] used a dataset twice as big as the DeepFashion2 training set, and the validation set contained another number of images. More details are in Table 1.

Metric	DeepMark	DF2 Paper
mAP_{box}	**0.723**	0.667
$mAP_{box}^{iou=0.50}$	**0.827**	0.814
$mAP_{box}^{iou=0.75}$	**0.795**	0.773

Table 3. Clothing detection results for DeepMark and Mask R-CNN[20]. The proposed model significantly outperforms Mask R-CNN[20] while been trained on half of the images.

NMS	-	✓	✓	✓
Flip	-	-	✓	✓
Multiscale	-	-	-	✓
mAP_{box}	0.707	0.710	0.723	**0.723**
mAP_{pt}	0.511	0.512	0.529	**0.532**
Inference time, ms	73	76	165	315

Table 4. Comparison of different post-processing strategies with Hourglass backbone.

	DLA-34		Hourglass	
	Single	Fusion	Single	Fusion
mAP_{box}	0.667	0.686	0.710	0.723
$mAP_{box}^{iou=0.50}$	0.788	0.802	0.821	0.827
$mAP_{box}^{iou=0.75}$	0.750	0.767	0.786	0.795
Inference time, ms	35	216	76	315

Table 5. Clothing detection.

	DLA-34		Hourglass	
	Single	Fusion	Single	Fusion
mAP_{pt}	0.469	0.513	0.582	0.601
	0.432	0.448	0.512	0.532
$mAP_{pt}^{oks=0.50}$	0.730	0.746	0.784	0.793
	0.716	0.730	0.774	0.784
$mAP_{pt}^{oks=0.75}$	0.547	0.563	0.660	0.683
	0.455	0.473	0.571	0.599
Inference time, ms	35	216	76	315

Table 6. Clothing landmark estimation.

best trade-off between speed and accuracy which makes it a good choice for the low-power devices. It works in only 35 ms per image on a single GPU, with an accuracy of 0.432 mAP_{pt} and 0.667 mAP_{box}. With the Hourglass model we have achieved the state-of-the-art results for both tasks.

5. Conclusion

The new approach, DeepMark, is proposed as an adaptation of CenterNet [1] for clothing landmark estimation tasks. The state-of-the-art accuracy was achieved on the DeepFashion2 dataset for 2 tasks: clothing detection (0.723 mAP) and clothing landmark estimation (0.532 mAP). The proposed neural network architecture has just one stage. It can be used without post-processing. It takes 35 ms per image and has quite high accuracy (0.432 mAP_{pt} and 0.667 mAP_{box}) for clothing detection tasks (Fig.1).

DeepMark has achieved the best results using the 8 GPU system with Nvidia GeForce RTX 2080ti graphics card after training the model for 63 epochs on the DeepFashion2 Challenge training set (191,961 images) with a batch of 37 images and input network resolution of 512×512. We used the CenterNet MS COCO model for object detection as the initial checkpoint and have performed experiments with two backbones: Hourglass and DLA-34. Learning rate schedule: 3e-4 - 50 epochs, 3e-5 - 3 epochs, 1e-5 - 10 epochs.

DeepMark outperformed the DeepFashion2 Challenge [2] baseline on the almost all metrics (Table 2). DeepFashion2 Challenge only provides a baseline for the landmark estimation problem, so comparison only with the paper [20] was provided for this task in Table 3.

Different inference strategies are considired in the paper including applying non-maximum suppression, fusing model outputs from original and flipped images with equal weighs and from original images; and images downscaled by 0.75. The results of these strategies for landmark estimation and object detection problems are shown in Table 4. The mixture of all inference strategies adds about 2 mAP over the vanilla solution.

Table 5 and Table 6 include results for object detection and landmark estimation problems for visible and visible+occluded landmarks. The DLA-34 model provides the

References

[1] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[2] Yuying Ge and Ping Luo. DeepFashion2 challenge, 2019.

[3] Syte. The leader in visual AI for retail, 2019.

[4] Vue.ai. Neural networks for retail, 2019.

[5] Oyper. We make all visual content shoppable, 2019.

[6] LOOKLET. Combining expertise from fashion, photography, and technology, 2019.

[7] WANNABY. Wanna kicks, 2019.

[8] GLAMSTORM. A virtual stylist as a bridge between the virtual and the real fashion world, 2017.

[9] Xianwang Wang and Tong Zhang. Clothes search in consumer photos via color matching and attribute learning. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1353–1356. ACM, 2011.

[10] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *European conference on computer vision*, pages 609–623. Springer, 2012.

[11] Shuai Zheng, Fan Yang, M Hadi Kiapour, and Robinson Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1670–1678. ACM, 2018.

[12] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. Hipster wars: Discovering elements of fashion styles. In *European conference on computer vision*, pages 472–488. Springer, 2014.

[13] 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.

[14] 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 *Proceedings of the IEEE international conference on computer vision*, pages 3343–3351, 2015.

[15] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5315–5324, 2015.

[16] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. Human parsing with contextualized convolutional neural network. In *Proceedings of the IEEE international conference on computer vision*, pages 1386–1394, 2015.

[17] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *European Conference on Computer Vision*, pages 229–245. Springer, 2016.

[18] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5334–5343, 2017.

[19] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE international conference on computer vision*, pages 1062–1070, 2015.

[20] 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 Conference on Computer Vision and Pattern Recognition*, pages 5337–5345, 2019.

[21] Deva Ramanan and Xiangxin Zhu. Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886. Citeseer, 2012.

[22] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807, 2014.
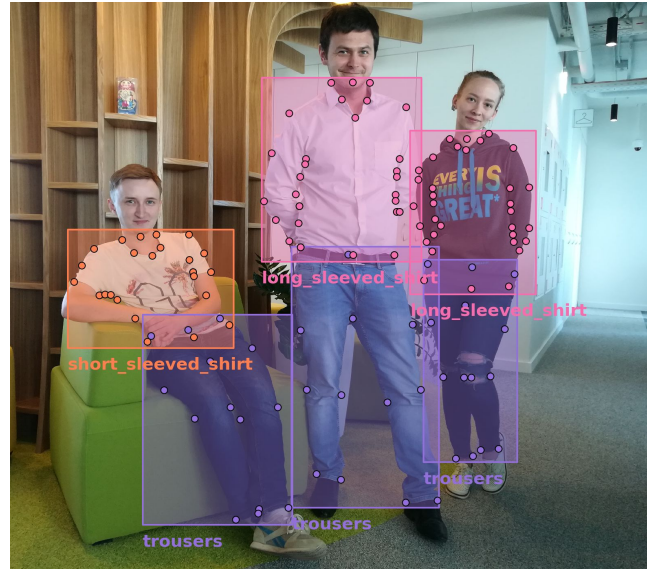
Figure 3. Example of DeepMark results. Best viewed in color.

[23] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[24] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 850–859, 2019.

[25] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[27] Alejandro Newell and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *CoRR*, abs/1611.05424, 2016.

[28] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. *CoRR*, abs/1803.08225, 2018.

[29] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. *CoRR*, abs/1312.4659, 2013.