

DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images

Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, Ping Luo

The Chinese University of Hong Kong

{yuyingge, ruimao.zhang}@cuhk.edu.hk, xgwang@ee.cuhk.edu.hk, {xtang, pluo}@ie.cuhk.edu.hk

Abstract

Understanding fashion images has been advanced by benchmarks with rich annotations such as DeepFashion, whose labels include clothing categories, landmarks, and consumer-commercial image pairs. However, DeepFashion has nonnegligible issues such as single clothing-item per image, sparse landmarks ($4 \sim 8$ only), and no per-pixel masks, making it had significant gap from real-world scenarios. We fill in the gap by presenting DeepFashion2 to address these issues. It is a versatile benchmark of four tasks including clothes detection, pose estimation, segmentation, and retrieval. It has 801K clothing items where each item has rich annotations such as style, scale, viewpoint, occlusion, bounding box, dense landmarks (e.g. 39 for ‘long sleeve outwear’ and 15 for ‘vest’), and masks. There are also 873K Commercial-Consumer clothes pairs. The annotations of DeepFashion2 are much larger than its counterparts such as 8 \times of FashionAI Global Challenge. A strong baseline is proposed, called Match R-CNN, which builds upon Mask R-CNN to solve the above four tasks in an end-to-end manner. Extensive evaluations are conducted with different criterions in DeepFashion2. DeepFashion2 Dataset will be released at : <https://github.com/switchablenorms/DeepFashion2>

1. Introduction

Fashion image analyses are active research topics in recent years because of their huge potential in industry. With the development of fashion datasets [20, 5, 7, 3, 14, 12, 21, 1], significant progresses have been achieved in this area [2, 19, 17, 18, 9, 8].

However, understanding fashion images remains a challenge in real-world applications, because of large deformations, occlusions, and discrepancies of clothes across domains between consumer and commercial images. Some



Figure 1. Comparisons between (a) DeepFashion and (b) DeepFashion2. (a) only has single item per image, which is annotated with $4 \sim 8$ sparse landmarks. The bounding boxes are estimated from the labeled landmarks, making them noisy. In (b), each image has minimum single item while maximum 7 items. Each item is manually labeled with bounding box, mask, dense landmarks (20 per item on average), and commercial-customer image pairs.

challenges can be rooted in the gap between the recent benchmark and the practical scenario. For example, the existing largest fashion dataset, DeepFashion [14], has its own drawbacks such as single clothing item per image, sparse landmark and pose definition (every clothing category shares the same definition of $4 \sim 8$ keypoints), and no per-pixel mask annotation as shown in Fig.1(a).

To address the above drawbacks, this work presents DeepFashion2, a large-scale benchmark with comprehensive tasks and annotations of fashion image understanding. DeepFashion2 contains 491K images of 13 popular clothing categories. A full spectrum of tasks are defined on them including clothes detection and recognition, landmark and pose estimation, segmentation, as well as verification and retrieval. All these tasks are supported by rich annotations.

tions. For instance, DeepFashion2 totally has 801K clothing items, where each item in an image is labeled with scale, occlusion, zooming, viewpoint, bounding box, dense landmarks, and per-pixel mask, as shown in Fig.1(b). These items can be grouped into 43.8K clothing identities, where a clothing identity represents the clothes that have almost the same cutting, pattern, and design. The images of the same identity are taken by both customers and commercial shopping stores. An item from the customer and an item from the commercial store forms a pair. There are 873K pairs that are 3.5 times larger than DeepFashion. The above thorough annotations enable developments of strong algorithms to understand fashion images.

This work has three main **contributions**. (1) We build a large-scale fashion benchmark with comprehensive tasks and annotations, to facilitate fashion image analysis. DeepFashion2 possesses the richest definitions of tasks and the largest number of labels. Its annotations are at least $3.5 \times$ of DeepFashion [14], $6.7 \times$ of ModaNet [21], and $8 \times$ of FashionAI [1]. (2) A full spectrum of tasks is carefully defined on the proposed dataset. For example, to our knowledge, clothing pose estimation is presented for the first time in the literature by defining landmarks and poses of 13 categories that are more diverse and fruitful than human pose. (3) With DeepFashion2, we extensively evaluate Mask R-CNN [6] that is a recent advanced framework for visual perception. A novel Match R-CNN is also proposed to aggregate all the learned features from clothes categories, poses, and masks to solve clothing image retrieval in an end-to-end manner. DeepFashion2 and implementations of Match R-CNN will be released.

1.1. Related Work

Clothes Datasets. Several clothes datasets have been proposed such as [20, 5, 7, 14, 21, 1] as summarized in Table 1. They vary in size as well as amount and type of annotations. For example, WTBI [5] and DARN [7] have 425K and 182K images respectively. They scraped category labels from metadata of the collected images from online shopping websites, making their labels noisy. In contrast, CCP [20], DeepFashion [14], and ModaNet [21] obtain category labels from human annotators. Moreover, different kinds of annotations are also provided in these datasets. For example, DeepFashion labels 4~8 landmarks (keypoints) per image that are defined on the functional regions of clothes (*e.g.* ‘collar’). The definitions of these sparse landmarks are shared across all categories, making them difficult to capture rich variations of clothing images. Furthermore, DeepFashion does not have mask annotations. By comparison, ModaNet [21] has street images with masks (polygons) of single person but without landmarks. Unlike existing datasets, DeepFashion2 contains 491K images and 801K instances of landmarks, masks, and bounding boxes,

	WTBI	DARN	DeepFashion	ModaNet	FashionAI	DeepFashion2
year	2015[5]	2015[7]	2016[14]	2018[21]	2018[1]	now
#images	425K	182K	800K	55K	357K	491K
#categories	11	20	50	13	41	13
#bboxe	39K	7K	×	×	×	801K
#landmarks	×	×	120K	×	100K	801K
#masks	×	×	×	119K	×	801K
#pairs	39K	91K	251K	×	×	873K

Table 1. Comparisons of DeepFashion2 with the other clothes datasets. The rows represent number of images, bounding boxes, landmarks, per-pixel masks, and consumer-to-shop pairs respectively. Bounding boxes inferred from other annotations are not counted.

as well as 873K pairs. It is the most comprehensive benchmark of its kinds to date.

Fashion Image Understanding. There are various tasks that analyze clothing images such as clothes detection [2, 14], landmark prediction [15, 19, 17], clothes segmentation [18, 20, 13], and retrieval [7, 5, 14]. However, a unify benchmark and framework to account for all these tasks is still desired. DeepFashion2 and Match R-CNN fill in this blank. We report extensive results for the above tasks with respect to different variations, including scale, occlusion, zoom-in, and viewpoint. For the task of clothes retrieval, unlike previous methods [5, 7] that performed image-level retrieval, DeepFashion2 enables instance-level retrieval of clothing items. We also present a new fashion task called clothes pose estimation, which is inspired by human pose estimation to predict clothing landmarks and skeletons for 13 clothes categories. This task helps improve performance of fashion image analysis in real-world applications.

2. DeepFashion2 Dataset and Benchmark

Overview. DeepFashion2 has four unique characteristics compared to existing fashion datasets. (1) *Large Sample Size.* It contains 491K images of 43.8K clothing identities of interest (unique garment displayed by shopping stores). On average, each identity has 12.7 items with different styles such as color and printing. DeepFashion2 contained 801K items in total. It is the largest fashion database to date. Furthermore, each item is associated with various annotations as introduced above.

(2) *Versatility.* DeepFashion2 is developed for multiple tasks of fashion understanding. Its rich annotations support clothes detection and classification, dense landmark and pose estimation, instance segmentation, and cross-domain instance-level clothes retrieval.

(3) *Expressivity.* This is mainly reflected in two aspects. First, multiple items are present in a single image, unlike DeepFashion where each image is labeled with at most one item. Second, we have 13 different definitions of landmarks and poses (skeletons) for 13 different categories. There is



Figure 2. Examples of DeepFashion2. The first column shows definitions of dense landmarks and skeletons of four categories. From (1) to (4), each row represents clothes images with different variations including ‘scale’, ‘occlusion’, ‘zoom-in’, and ‘viewpoint’. At each row, we partition the images into two groups, the left three columns represent clothes from commercial stores, while the right three columns are from customers. In each group, the three images indicate three levels of difficulty with respect to the corresponding variation, including (1) ‘small’, ‘moderate’, ‘large’ scale, (2) ‘slight’, ‘medium’, ‘heavy’ occlusion, (3) ‘no’, ‘medium’, ‘large’ zoom-in, (4) ‘not on human’, ‘side’, ‘back’ viewpoint. Furthermore, at each row, the items in these two groups of images are from the same clothing identity but from two different domains, that is, commercial and customer. The items of the same identity may have different styles such as color and printing. Each item is annotated with landmarks and masks.

23 defined landmarks for each category on average. Some definitions are shown in the first column of Fig.2. These representations are different from human pose and are not presented in previous work. They facilitate learning of strong clothes features that satisfy real-world requirements.

(4) **Diversity.** We collect data by controlling their variations in terms of four properties including scale, occlusion, zoom-in, and viewpoint as illustrated in Fig.2, making DeepFashion2 a challenging benchmark. For each property, each clothing item is assigned to one of three levels of difficulty. Fig.2 shows that each identity has high diversity where its items are from different difficulties.

Data Collection and Cleaning. Raw data of DeepFashion2 are collected from two sources including DeepFashion [14] and online shopping websites. In particular, images of each consumer-to-shop pair in DeepFashion are included in DeepFashion2, while the other images are removed. We

further crawl a large set of images on the Internet from both commercial shopping stores and consumers. To clean up the crawled set, we first remove shop images with no corresponding consumer-taken photos. Then human annotators are asked to clean images that contain clothes with large occlusions, small scales, and low resolutions. Eventually we have 491K images of 801K items and 873K commercial-consumer pairs.

Variations. We explain the variations in DeepFashion2. Their statistics are plotted in Fig.3. (1) **Scale.** We divide all clothing items into three sets, according to the proportion of an item compared to the image size, including ‘small’ ($< 10\%$), ‘moderate’ ($10\% \sim 40\%$), and ‘large’ ($> 40\%$). Fig.3(a) shows that only 50% items have moderate scale. (2) **Occlusion.** An item with occlusion means that its region is occluded by hair, human body, accessory or other items. Note that an item with its region outside the im-

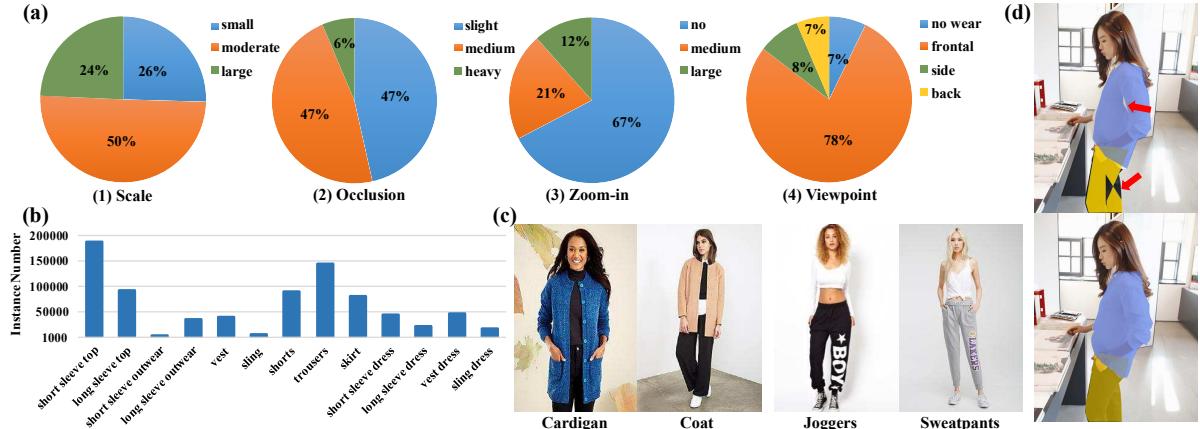


Figure 3. (a) shows the statistics of different variations in DeepFashion2. (b) is the numbers of items of the 13 categories in DeepFashion2. (c) shows that categories in DeepFashion [14] have ambiguity. For example, it is difficult to distinguish between ‘cardigan’ and ‘coat’, and between ‘joggers’ and ‘sweatpants’. They result in ambiguity when labeling data. (d) **Top:** masks may be inaccurate when complex poses are presented. **Bottom:** the masks will be refined by human.

age does not belong to this case. Each item is categorized by the number of its landmarks that are occluded, including ‘partial occlusion’ (< 20% occluded keypoints), ‘heavy occlusion’ (> 50% occluded keypoints), ‘medium occlusion’ (otherwise). More than 50% items have medium or heavy occlusions as summarized in Fig.3. (3) **Zoom-in**. An item with zoom-in means that its region is outside the image. This is categorized by the number of landmarks outside image. We define ‘no’, ‘large’ (> 30%), and ‘medium’ zoom-in. We see that more than 30% items are zoomed in. (4) **Viewpoint**. We divide all items into four partitions including 7% clothes that are not on people, 78% clothes on people from frontal viewpoint, 15% clothes on people from side or back viewpoint.

2.1. Data Labeling

Category and Bounding Box. Human annotators are asked to draw a bounding box and assign a category label for each clothing item. DeepFashion [14] defines 50 categories but half of them contain less than 5% number of images. Also, ambiguity exists between 50 categories making data labeling difficult as shown in Fig.3(c). By grouping categories in DeepFashion, we derive 13 popular categories without ambiguity. The numbers of items of 13 categories are shown in Fig.3(b).

Clothes Landmark, Contour, and Skeleton. As different categories of clothes (*e.g.* upper- and lower-body garment) have different deformations and appearance changes, we represent each category by defining its pose, which is a set of landmarks as well as contours and skeletons between landmarks. They capture shapes and structures of clothes. Pose definitions are not presented in previous work and are significantly different from human pose. For each clothing item of a category, human annotations are asked to label

landmarks following these instructions.

Moreover, each landmark is assigned one of the two modes, ‘visible’ or ‘occluded’. We then generate contours and skeletons automatically by connecting landmarks in a certain order. To facilitate this process, annotators are also asked to distinguish landmarks into two types, that is, contour point or junction point. The former one refers to keypoints at the boundary of an item, while the latter one is assigned to keypoints in conjunction *e.g.* ‘endpoint of strap on sling’. The above process controls the labeling quality, because the generated skeletons help the annotators reexamine whether the landmarks are labeled with good quality. In particular, only when the contour covers the entire item, the labeled results are eligible, otherwise keypoints will be refined.

Mask. We label per-pixel mask for each item in a semi-automatic manner with two stages. The first stage automatically generates masks from the contours. In the second stage, human annotators are asked to refine the masks, because the generated masks may be not accurate when complex human poses are presented. As shown in Fig.3(d), the mark is inaccurate when an image is taken from side-view of people crossing legs. The masks will be refined by human.

Style. As introduced before, we collect 43.8K different clothing identities where each identity has 13 items on average. These items are further labeled with different styles such as color, printing, and logo. Fig.2 shows that a pair of clothes that have the same identity could have different styles.

2.2. Benchmarks

We build four benchmarks by using the images and labels from DeepFashion2. For each benchmark, there are

391K images for training, 34K images for validation and 67K images for test.

Clothes Detection. This task detects clothes in an image by predicting bounding boxes and category labels. The evaluation metrics are the bounding box’s average precision AP_{box} , $AP_{box}^{IoU=0.50}$, and $AP_{box}^{IoU=0.75}$ by following COCO [11].

Landmark Estimation. This task aims to predict landmarks for each detected clothing item in an each image. Similarly, we employ the evaluation metrics used by COCO for human pose estimation by calculating the average precision for keypoints AP_{pt} , $AP_{pt}^{OKS=0.50}$, and $AP_{pt}^{OKS=0.75}$, where OKS indicates the object landmark similarity.

Segmentation. This task assigns a category label (including background label) to each pixel in an item. The evaluation metrics is the average precision including AP_{mask} , $AP_{mask}^{IoU=0.50}$, and $AP_{mask}^{IoU=0.75}$ computed over masks.

Commercial-Consumer Clothes Retrieval. Given a detected item from a consumer-taken photo, this task aims to search the commercial images in the gallery for the items that are corresponding to this detected item. This setting is more realistic than DeepFashion [14], which assumes ground-truth bounding box is provided. In this task, top-k retrieval accuracy is employed as the evaluation metric. We emphasize the retrieval performance while still consider the influence of detector. If a clothing item fails to be detected, this query item is counted as missed. In particular, we have more than 686K commercial-consumer clothes pairs in the training set. In the validation set, there are 10,990 consumer images with 12,550 items as a query set, and 21,438 commercial images with 37,183 items as a gallery set. In the test set, there are 21,550 consumer images with 24,402 items as queries, while 43,608 commercial images with 75,347 items in the gallery.

3. Match R-CNN

We present a strong baseline model built upon Mask R-CNN [6] for DeepFashion2, termed Match R-CNN, which is an end-to-end training framework that jointly learns clothes detection, landmark estimation, instance segmentation, and consumer-to-shop retrieval. The above tasks are solved by using different streams and stacking a Siamese module on top of these streams to aggregate learned features.

As shown in Fig.4, Match R-CNN employs two images I_1 and I_2 as inputs. Each image is passed through three main components including a Feature Network (FN), a Perception Network (PN), and a Matching Network (MN). In the first stage, FN contains a ResNet-FPN [10] backbone, a region proposal network (RPN) [16] and RoIAlign module. An image is first fed into ResNet50 to extract features, which are then fed into a FPN that uses a top-down architec-

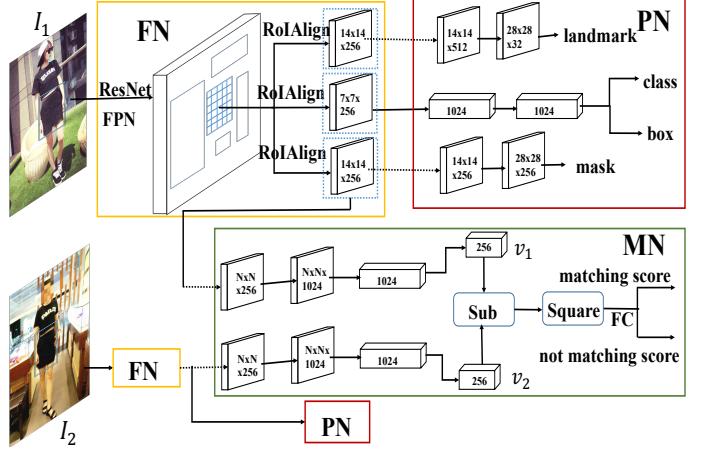


Figure 4. **Diagram of Match R-CNN** that contains three main components including a feature extraction network (FN), a perception network (PN), and a match network (MN).

ture with lateral connections to build a pyramid of feature maps. RoIAlign extracts features from different levels of the pyramid map.

In the second stage, PN contains three streams of networks including landmark estimation, clothes detection, and mask prediction as shown in Fig.4. The extracted RoI features after the first stage are fed into three streams in PN separately. The clothes detection stream has two hidden fully-connected (fc) layers, one fc layer for classification, and one fc layer for bounding box regression. The stream of landmark estimation has 8 ‘conv’ layers and 2 ‘deconv’ layers to predict landmarks. Segmentation stream has 4 ‘conv’ layers, 1 ‘deconv’ layer, and another ‘conv’ layer to predict masks.

In the third stage, MN contains a feature extractor and a similarity learning network for clothes retrieval. The learned RoI features after the FN component are highly discriminative with respect to clothes category, pose, and mask. They are fed into MN to obtain features vectors for retrieval, where v_1 and v_2 are passed into the similarity learning network to obtain the similarity score between the detected clothing items in I_1 and I_2 . Specifically, the feature extractor has 4 ‘conv’ layers, one pooling layer, and one fc layer. The similarity learning network consists of subtraction and square operator and a fc layer, which estimates the probability of whether two clothing items match or not.

Loss Functions. The parameters Θ of the Match R-CNN are optimized by minimizing five loss functions, which are formulated as $\min_{\Theta} \mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{box} + \lambda_3 \mathcal{L}_{pose} + \lambda_4 \mathcal{L}_{mask} + \lambda_5 \mathcal{L}_{pair}$, including a cross-entropy (CE) loss \mathcal{L}_{cls} for clothes classification, a smooth loss [4] \mathcal{L}_{box} for bounding box regression, a CE loss \mathcal{L}_{pose} for landmark es-

	scale	small	moderate	large	occlusion	slight	medium	heavy	zoom-in	no	medium	large	viewpoint	frontal	side or back	overall
AP _{box}		0.604	0.700	0.660	0.712	0.654	0.372	0.695	0.629	0.466	0.624	0.681	0.641	0.667		
AP _{box} ^{IoU=0.50}		0.780	0.851	0.768	0.844	0.810	0.531	0.848	0.755	0.563	0.713	0.832	0.796	0.814		
AP _{box} ^{IoU=0.75}		0.717	0.809	0.744	0.812	0.768	0.433	0.806	0.718	0.525	0.688	0.791	0.744	0.773		

Table 2. Clothes detection of Mask R-CNN [6] on different validation subsets, including scale, occlusion, zoom-in, and viewpoint. The evaluation metrics are AP_{box}, AP_{box}^{IoU=0.50}, and AP_{box}^{IoU=0.75}. The best performance of each subset is bold.

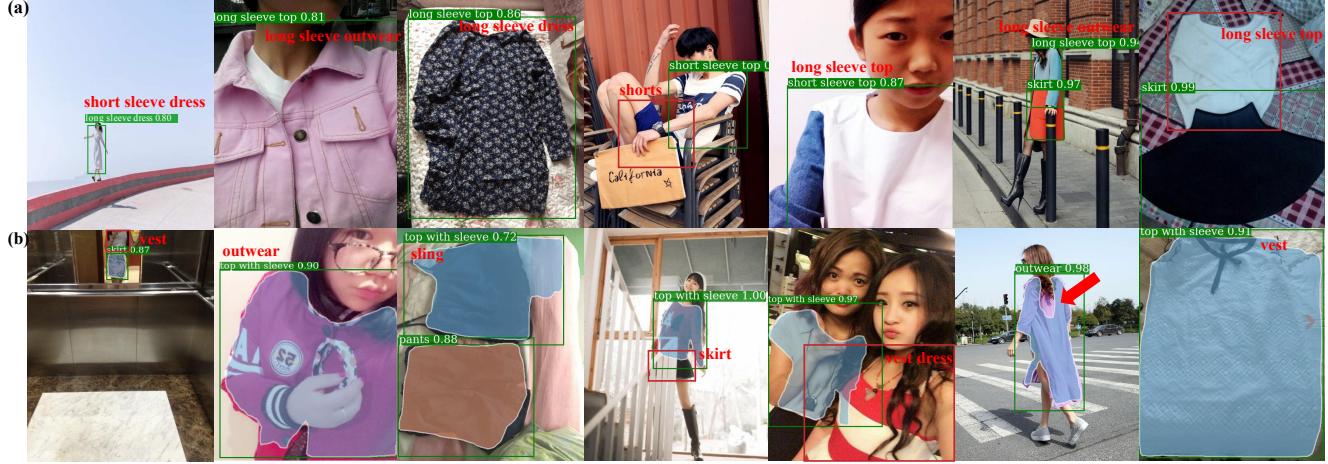


Figure 5. (a) shows failure cases in clothes detection while (b) shows failure cases in clothes segmentation. In (a) and (b), the missing bounding boxes are drawn in red while the correct category labels are also in red. Inaccurate masks are also highlighted by arrows in (b). For example, clothes fail to be detected or segmented in too small scale, too large scale, large non-rigid deformation, heavy occlusion, large zoom-in, side or back viewpoint.

timation, a CE loss \mathcal{L}_{mask} for clothes segmentation, and a CE loss \mathcal{L}_{pair} for clothes retrieval. Specifically, \mathcal{L}_{cls} , \mathcal{L}_{box} , \mathcal{L}_{pose} , and \mathcal{L}_{mask} are identical as defined in [6]. We have $\mathcal{L}_{pair} = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$, where $y_i = 1$ indicates the two items of a pair are matched, otherwise $y_i = 0$.

Implementations. In our experiments, each training image is resized to its shorter edge of 800 pixels with its longer edge that is no more than 1333 pixels. Each minibatch has two images in a GPU and 8 GPUs are used for training. For minibatch size 16, the learning rate (LR) schedule starts at 0.02 and is decreased by a factor of 0.1 after 8 epochs and then 11 epochs, and finally terminates at 12 epochs. This scheduler is denoted as 1x. Mask R-CNN adopts 2x schedule for clothes detection and segmentation where ‘2x’ is twice as long as 1x with the LR scaled proportionally. Then It adopts s1x for landmark and pose estimation where s1x scales the 1x schedule by roughly 1.44x. Match R-CNN uses 1x schedule for consumer-to-shop clothes retrieval. The above models are trained by using SGD.

Inference. At testing time, images are resized in the same way as the training stage. The top 1000 proposals with detection probabilities are chosen for bounding box classification and regression. Then non-maximum suppression is applied to these proposals. The filtered proposals are fed

into the landmark branch and the mask branch separately. For the retrieval task, each unique detected clothing item in consumer-taken image with highest confidence is selected as query.

4. Experiments

We demonstrate the effectiveness of DeepFashion2 by evaluating Mask R-CNN [6] and Match R-CNN in multiple tasks including clothes detection and classification, landmark estimation, instance segmentation, and consumer-to-shop clothes retrieval. To further show the large variations of DeepFashion2, the validation set is divided into three subsets according to their difficulty levels in scale, occlusion, zoom-in, and viewpoint. The settings of Mask R-CNN and Match R-CNN follow Sec.3.

The following sections from 4.1 to 4.4 report results for different tasks, showing that DeepFashion2 imposes significant challenges to both Mask R-CNN and Match R-CNN, which are the recent state-of-the-art systems for visual perception.

4.1. Clothes Detection

Table 2 summarizes the results of clothes detection on different difficulty subsets. We see that the clothes of mod-

	scale			occlusion			zoom-in			viewpoint			overall
	small	moderate	large	slight	medium	heavy	no	medium	large	no wear	frontal	side or back	
AP _{pt}	0.587	0.687	0.599	0.669	0.631	0.398	0.688	0.559	0.375	0.527	0.677	0.536	0.641
	0.497	0.607	0.555	0.643	0.530	0.248	0.616	0.489	0.319	0.510	0.596	0.456	0.563
AP _{pt} ^{OKS=0.50}	0.780	0.854	0.782	0.851	0.813	0.534	0.855	0.757	0.571	0.724	0.846	0.748	0.820
	0.764	0.839	0.774	0.847	0.799	0.479	0.848	0.744	0.549	0.716	0.832	0.727	0.805
AP _{pt} ^{OKS=0.75}	0.671	0.779	0.678	0.760	0.718	0.440	0.786	0.633	0.390	0.571	0.771	0.610	0.728
	0.551	0.703	0.625	0.739	0.600	0.236	0.714	0.537	0.307	0.550	0.684	0.506	0.641

Table 3. **Landmark estimation** of Mask R-CNN [6] on different validation subsets, including scale, occlusion, zoom-in, and viewpoint. Results of evaluation on visible landmarks only and evaluation on both visible and occlusion landmarks are separately shown in each row. The evaluation metrics are AP_{pt}, AP_{pt}^{OKS=0.50}, and AP_{pt}^{OKS=0.75}. The best performance of each subset is bold.

	scale			occlusion			zoom-in			viewpoint			overall
	small	moderate	large	slight	medium	heavy	no	medium	large	no wear	frontal	side or back	
AP _{mask}	0.634	0.703	0.666	0.720	0.656	0.381	0.701	0.637	0.478	0.664	0.689	0.635	0.674
AP _{mask} ^{IoU=0.50}	0.811	0.865	0.798	0.863	0.824	0.543	0.861	0.791	0.591	0.757	0.849	0.811	0.834
AP _{mask} ^{IoU=0.75}	0.752	0.826	0.773	0.836	0.780	0.444	0.823	0.751	0.559	0.737	0.810	0.755	0.793

Table 4. **Clothes segmentation** of Mask R-CNN [6] on different validation subsets, including scale, occlusion, zoom-in, and viewpoint. The evaluation metrics are AP_{mask}, AP_{mask}^{IoU=0.50}, and AP_{mask}^{IoU=0.75}. The best performance of each subset is bold.

erate scale, slight occlusion, no zoom-in, and frontal viewpoint have the highest detection rates. There are several observations. First, detecting clothes with small or large scale reduces detection rates. Some failure cases are provided in Fig.5(a) where the item could occupy less than 2% of the image while some occupies more than 90% of the image. Second, in Table 2, it is intuitively to see that heavy occlusion and large zoom-in degenerate performance. In these two cases, large portions of the clothes are invisible as shown in Fig.5(a). Third, it is seen in Table 2 that the clothing items not on human body also drop performance. This is because they possess large non-rigid deformations as visualized in the failure cases of Fig.5(a). These variations are not presented in previous object detection benchmarks such as COCO. Fourth, clothes with side or back viewpoint, are much more difficult to detect as shown in Fig.5(a).

4.2. Landmark and Pose Estimation

Table 3 summarizes the results of landmark estimation. The evaluation of each subset is performed in two settings, including visible landmark only (the occluded landmarks are not evaluated), as well as both visible and occluded landmarks. As estimating the occluded landmarks is more difficult than visible landmarks, the second setting generally provides worse results than the first setting.

In general, we see that Mask R-CNN obtains an overall AP of just 0.563, showing that clothes landmark estimation could be even more challenging than human pose estimation in COCO. In particular, Table 3 exhibits similar trends as those from clothes detection. For example, the clothing items with moderate scale, slight occlusion, no zoom-in, and frontal viewpoint have better results than the others subsets. Moreover, heavy occlusion and zoom-in decreases performance a lot. Some results are given in Fig.6(a).

4.3. Clothes Segmentation

Table 4 summarizes the results of segmentation. The performance declines when segmenting clothing items with small and large scale, heavy occlusion, large zoom-in, side or back viewpoint, which is consistent with those trends in the previous tasks. Some results are given in Fig.6(b). Some failure cases are visualized in Fig.5(b).

4.4. Consumer-to-Shop Clothes Retrieval

Table 5 summarizes the results of clothes retrieval. The retrieval accuracy is reported in Fig. 6(d), where top-1, -5, -10, and -20 retrieval accuracy are shown. We evaluate two settings in (c.1) and (c.2), when the bounding boxes are predicted by the detection module in Match R-CNN and are provided as ground truths. Match R-CNN achieves a top-20 accuracy of less than 0.7 with ground-truth bounding boxes provided, indicating that the retrieval benchmark is challenging. Furthermore, retrieval accuracy drops when using detected boxes, meaning that this is a more realistic setting.

In Table 5, different combinations of the learned features are also evaluated. In general, the combination of features increases the accuracy. In particular, the learned features from pose and class achieve better results than the other features. When comparing learned features from pose and mask, we find that the former achieves better results, indicating that landmark locations can be more robust across scenarios.

As shown in Table 5, the performance declines when small scale, heavily occluded clothing items are presented. Clothes with large zoom-in achieved the lowest accuracy because only part of clothes are displayed in the image and crucial distinguishable features may be missing. Compared

	scale	occlusion	zoom-in	viewpoint	overall							
	small	moderate	large	no	medium	large	no wear	frontal	side or back	top-1	top-10	top-20
class	0.520	0.630	0.540	0.572	0.563	0.558	0.618	0.547	0.444	0.546	0.584	0.533
	0.485	0.537	0.502	0.527	0.508	0.383	0.553	0.496	0.405	0.499	0.523	0.487
pose	0.721	0.778	0.735	0.756	0.737	0.728	0.775	0.751	0.621	0.731	0.763	0.711
	0.637	0.702	0.691	0.710	0.670	0.580	0.710	0.701	0.560	0.690	0.700	0.645
mask	0.624	0.714	0.646	0.675	0.651	0.632	0.711	0.655	0.526	0.644	0.682	0.637
	0.552	0.657	0.608	0.639	0.593	0.555	0.654	0.613	0.495	0.615	0.630	0.565
pose+class	0.752	0.786	0.733	0.754	0.750	0.728	0.789	0.750	0.620	0.726	0.771	0.719
	0.691	0.730	0.705	0.725	0.706	0.605	0.746	0.709	0.582	0.699	0.723	0.684
mask+class	0.656	0.728	0.687	0.714	0.676	0.654	0.725	0.702	0.565	0.684	0.712	0.658
	0.610	0.666	0.649	0.676	0.623	0.549	0.674	0.655	0.536	0.648	0.661	0.604

Table 5. **Consumer-to-Shop Clothes Retrieval** of Match R-CNN on different subsets of some validation consumer-taken images. Each query item in these images has over 5 identical clothing items in validation commercial images. Results of evaluation on ground truth box and detected box are separately shown in each row. The evaluation metrics are top-20 accuracy. The best performance of each subset is bold.



Figure 6. (a) shows results of landmark and pose estimation. (b) shows results of clothes segmentation. (c) shows queries with top-5 retrieved clothing items. The first column is the image from the customer with bounding box predicted by detection module, and the second to the sixth columns show the retrieval results from the store. (d) is the retrieval accuracy of overall query validation set with (1) detected box (2) ground truth box. Evaluation metrics are top-1, -5, -10, -15, and -20 retrieval accuracy.

with clothes on people from frontal view, clothes from side or back viewpoint perform worse due to lack of discriminative features like patterns on the front of tops. Example queries with top-5 retrieved clothing items are shown in Fig.6(c).

5. Conclusions

This work represented DeepFashion2, a large-scale fashion image benchmark with comprehensive tasks and annotations. DeepFashion2 contains 491K images, each of which is richly labeled with style, scale, occlusion, zooming, viewpoint, bounding box, dense landmarks and pose, pixel-level masks, and pair of images of identical item from consumer and commercial store. We establish benchmarks covering multiple tasks in fashion understanding, including clothes detection, landmark and pose estimation, clothes segmentation, consumer-to-shop verification and retrieval. A novel Match R-CNN framework that builds upon Mask R-CNN is proposed to solve the above tasks in end-to-end manner. Extensive evaluations are conducted in DeepFashion2.

The rich data and labels of DeepFashion2 will definitely facilitate the developments of algorithms to understand fashion images in future work. We will focus on three aspects. First, more challenging tasks will be explored with DeepFashion2, such as synthesizing clothing images by using GANs. Second, it is also interesting to explore multi-domain learning for clothing images, because fashion trends of clothes may change frequently, making variations of clothing images changed. Third, we will introduce more evaluation metrics into DeepFashion2, such as size, runtime, and memory consumptions of deep models, towards understanding fashion images in real-world scenario.

Acknowledgement This work is supported in part by SenseTime Group Limited, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants CUHK14202217, CUHK14203118, CUHK14205615, CUHK14207814, CUHK14213616.

References

- [1] Fashionai dataset. <http://fashionai.alibaba.com/datasets/>.
- [2] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *ECCV*, 2012.
- [3] 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 *CVPR*, 2015.
- [4] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.
- [5] 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 *ICCV*, 2015.
- [6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.
- [7] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015.
- [8] Xin Ji, Wei Wang, Meihui Zhang, and Yang Yang. Cross-domain image retrieval with attention modeling. In *ACM Multimedia*, 2017.
- [9] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. Interpretable multimodal retrieval for fashion products. In *ACM Multimedia*, 2018.
- [10] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.
- [11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [12] Kuan-Hsien Liu, Ting-Yen Chen, and Chu-Song Chen. Mvc: A dataset for view-invariant clothing retrieval and attribute prediction. In *ACM Multimedia*, 2016.
- [13] Si Liu, Xiaodan Liang, Luoqi Liu, Ke Lu, Liang Lin, Xiaochun Cao, and Shuicheng Yan. Fashion parsing with video context. *IEEE Transactions on Multimedia*, 17(8):1347–1358, 2015.
- [14] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.
- [15] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *ECCV*, 2016.
- [16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.
- [17] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, 2018.
- [18] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *ICCV*, 2013.
- [19] Sijie Yan, Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *ACM Multimedia*, 2017.
- [20] Wei Yang, Ping Luo, and Liang Lin. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*, 2014.
- [21] Shuai Zheng, Fan Yang, M Hadi Kiapour, and Robinson Piрамуту. Modanet: A large-scale street fashion dataset with polygon annotations. In *ACM Multimedia*, 2018.