Fashionpedia: Ontology, Segmentation, and an Attribute Localization Dataset

Menglin Jia*[1], Mengyun Shi*[1,4], Mikhail Sirotenko*[3], Yin Cui*[3],
Claire Cardie[1], Bharath Hariharan[1], Hartwig Adam[3], Serge Belongie[1,2]

[1]Cornell University [2]Cornell Tech [3]Google Research [4]Hearst Magazines

Abstract. In this work we explore the task of *instance segmentation with attribute localization*, which unifies instance segmentation (detect and segment each object instance) and fine-grained visual attribute categorization (recognize one or multiple attributes). The proposed task requires both localizing an object and describing its properties. To illustrate the various aspects of this task, we focus on the domain of fashion and introduce *Fashionpedia* as a step toward mapping out the visual aspects of the fashion world. Fashionpedia consists of two parts: (1) an ontology built by fashion experts containing 27 main apparel categories, 19 apparel parts, 294 fine-grained attributes and their relationships; (2) a dataset with everyday and celebrity event fashion images annotated with segmentation masks and their associated per-mask fine-grained attributes, built upon the Fashionpedia ontology. In order to solve this challenging task, we propose a novel Attribute-Mask R-CNN model to jointly perform instance segmentation and localized attribute recognition, and provide a novel evaluation metric for the task. Fashionpedia is available at: https://fashionpedia.github.io/home/.

Keywords: Dataset, Ontology, Instance Segmentation, Fine-Grained, Attribute, Fashion

1 Introduction

Recent progress in the field of computer vision has advanced machines' ability to recognize and understand our visual world, showing significant impacts in fields including autonomous driving [52], product recognition [32,14], *etc*. These real-world applications are fueled by various visual understanding tasks with the goals of *naming*, *describing (attribute recognition)*, or *localizing* objects within an image.

Naming and localizing objects is formulated as an object detection task (Figure 1(a-c)). As a hallmark for computer recognition, this task is to identify and indicate the boundaries of objects in the form of bounding boxes or segmentation masks [12,37,17]. *Attribute recognition* [6,29,9,36] (Figure 1(d)) instead focuses

* equal contribution.

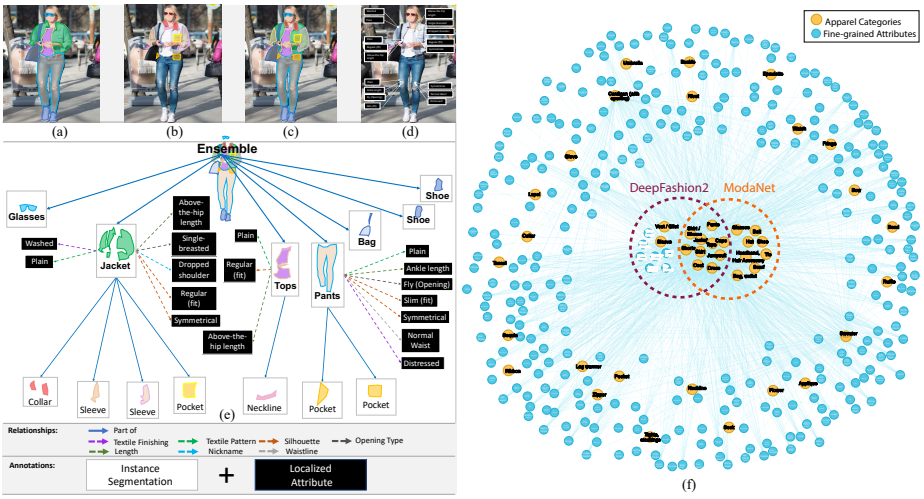


Fig. 1: **An illustration of the Fashionpedia dataset and ontology:** (a) main garment masks; (b) garment part masks; (c) both main garment and garment part masks; (d) fine-grained apparel attributes; (e) an exploded view of the annotation diagram: the image is annotated with both instance segmentation masks *(white boxes)* and per-mask fine-grained attributes *(black boxes)*; (f) visualization of the Fashionpedia ontology: we created Fashionpedia ontology and separate the concept of categories *(yellow nodes)* and attributes [38] *(blue nodes)* in fashion. It covers pre-defined garment categories used by both Deepfashion2 [11] and ModaNet [53]. Mapping with DeepFashion2 also shows the versatility of using attributes and categories. We are able to present all 13 garment classes in DeepFashion2 with 11 main garment categories, 1 garment part, and 7 attributes

on describing and comparing objects, since an object also has many other properties or attributes in addition to its category. Attributes not only provide a compact and scalable way to represent objects in the world, as pointed out by Ferrari and Zisserman [9], attribute learning also enables transfer of existing knowledge to novel classes. This is particularly useful for fine-grained visual recognition, with the goal of distinguishing subordinate visual categories such as birds [46] or natural species [43].

In the spirit of mapping the visual world, we propose a new task, *instance segmentation with attribute localization*, which unifies object detection and fine-grained attribute recognition. As illustrated in Figure 1(e), this task offers a structured representation of an image. Automatic recognition of a rich set of attributes for each segmented object instance complements category-level object detection and therefore advance the degree of complexity of images and scenes we can make understandable to machines. In this work, we focus on the fashion domain as an example to illustrate this task. Fashion comes with rich and

complex apparel with attributes, influences many aspects of modern societies, and has a strong financial and cultural impact. We anticipate that the proposed task is also suitable for other man-made product domains such as automobile and home interior.

Structured representations of images often rely on structured vocabularies [28]. With this in mind, we construct the Fashionpedia ontology (Figure 1(f)) and image dataset (Figure 1(a-e)), annotating fashion images with detailed segmentation masks for apparel categories, parts, and their attributes. Our proposed ontology provides a rich schema for interpretation and organization of individuals' garments, styles, or fashion collections [24]. For example, we can create a knowledge graph (see supplementary material for more details) by aggregating structured information within each image and exploiting relationships between garments and garment parts, categories, and attributes in the Fashionpedia ontology. Our insight is that a large-scale fashion segmentation and attribute localization dataset built with a fashion ontology can help computer vision models achieve better performance on fine-grained image understanding and reasoning tasks.

The contributions of this work are as follows:

A *novel task* of fine-grained instance segmentation with attribute localization. The proposed task unifies instance segmentation and visual attribute recognition, which is an important step toward structural understanding of visual content in real-world applications.

A *unified fashion ontology* informed by product descriptions from the internet and built by fashion experts. Our ontology captures the complex structure of fashion objects and ambiguity in descriptions obtained from the web, containing 46 apparel objects (27 main apparels and 19 apparel parts), and 294 fine-grained attributes (spanning 9 super categories) in total. To facilitate the development of related efforts, we also provide a mapping with categories from existing fashion segmentation datasets, see Figure 1(f).

A *dataset* with a total of 48,825 clothing images in daily-life, street-style, celebrity events, runway, and online shopping annotated both by crowd workers for segmentation masks and fashion experts for localized attributes, with the goal of developing and benchmarking computer vision models for comprehensive understanding of fashion.

A new *model*, Attribute-Mask R-CNN, is proposed with the aim of jointly performing instance segmentation and localized attribute recognition; a novel *evaluation metric* for this task is also provided.

2 Related Work

The combined task of fine-grained instance segmentation and attribute localization has not received a great deal of attention in the literature. On one hand, COCO [31] and LVIS [15] represent the benchmarks of object detection for common objects. Panoptic segmentation is proposed to unify both semantic and instance segmentation, addressing both stuff and thing classes [27]. In

Table 1: Comparison of fashion-related datasets: (Cls. = Classification, Segm. = Segmentation, MG = Main Garment, GP = Garment Part, A = Accessory, S = Style, FGC = Fine-Grained Categorization). To the best of our knowledge, we include all fashion-related datasets focusing on visual recognition

Name	**Category Annotation Type**			**Attribute Annotation Type**		
	Cls.	**BBox**	**Segm.**	**Unlocalized**	**Localized**	**FGC**
Clothing Parsing [50]	MG, A	-	-	-	-	-
Chic or Social [49]	MG, A	-	-	-	-	-
Hipster [26]	MG, A, S	-	-	-	-	-
Ups and Downs [19]	MG	-	-	-	-	-
Fashion550k [23]	MG, A	-	-	-	-	-
Fashion-MNIST [48]	MG	-	-	-	-	-
Runway2Realway [44]	-	-	MG, A	-	-	-
ModaNet [53]	-	MG, A	MG, A	-	-	-
Deepfashion2 [11]	-	MG	MG	-	-	-
Fashion144k [40]	MG, A	-	-	✓	-	-
Fashion Style-128 Floats [41]	S	-	-	✓	-	-
UT Zappos50K [51]	A	-	-	✓	-	-
Fashion200K [16]	MG	-	-	✓	-	-
FashionStyle14 [42]	S	-	-	✓	-	-
Main Product Detection [39]	-	MG	-	✓	-	-
StreetStyle-27K [35]	-	-	-	✓	-	✓
UT-latent look [21]	MG, S	-	-	✓	-	✓
FashionAI [7]	MG, GP, A	-	-	✓	-	✓
iMat-Fashion Attribute [14]	MG, GP, A, S	-	-	✓	-	✓
Apparel classification-Style [4]	-	MG	-	✓	-	✓
DARN [22]	-	MG	-	✓	-	✓
WTBI [25]	-	MG, A	-	✓	-	✓
Deepfashion [32]	S	MG	-	✓	-	✓
Fashionpedia	-	**MG, GP, A**	**MG, GP, A**	-	✓	✓

spite of the domain differences, Fashionpedia has comparable mask qualities with LVIS and the similar total number of segmentation masks as COCO. On the other hand, we have also observed an increasing effort to curate datasets for fine-grained visual recognition, evolved from CUB-200 Birds [46] to the recent iNaturalist dataset [43]. The goal of this line of work is to advance the state-of-the-art in automatic image classification for large numbers of real world, fine-grained categories. A rather unexplored area of these datasets, however, is to provide a structured representation of an image. Visual Genome [28] provides dense annotations of object bounding boxes, attributes, and relationships in the general domain, enabling a structured representation of the image. In our work, we instead focus on fine-grained attributes and provide segmentation masks in the fashion domain to advance the clothing recognition task.

Clothing recognition has received increasing attention in the computer vision community recently. A number of works provide valuable apparel-related datasets [50,4,49,26,44,40,22,25,19,41,32,23,48,51,16,42,39,35,21,53,7,11]. These pioneering works enabled several recent advances in clothing-related recognition

and knowledge discovery [10,34]. Table 1 summarizes the comparison among different fashion datasets regarding annotation types of clothing categories and attributes. Our dataset distinguishes itself in the following three aspects.

Exhaustive annotation of segmentation masks: Existing fashion datasets [44,53,11] offer segmentation masks for the main garment (e.g., jacket, coat, dress) and the accessory categories (e.g., bag, shoe). Smaller garment objects such as collars and pockets are not annotated. However, these small objects could be valuable for real-world applications (e.g., searching for a specific collar shape during online-shopping). Our dataset is not only annotated with the segmentation masks for a total of 27 main garments and accessory categories but also 19 garment parts (e.g., collar, sleeve, pocket, zipper, embroidery).

Localized attributes: The fine-grained attributes from existing datasets [22,32,39,14] tend to be noisy, mainly because most of the annotations are collected by crawling fashion product attribute-level descriptions directly from large online shopping websites. Unlike these datasets, fine-grained attributes in our dataset are annotated manually by fashion experts. To the best of our knowledge, ours is the only dataset to annotate localized attributes: fashion experts are asked to annotate attributes associated with the segmentation masks labeled by crowd workers. Localized attributes could potentially help computational models detect and understand attributes more accurately.

Fine-grained categorization: Previous studies on fine-grained attribute categorization suffer from several issues including: (1) repeated attributes belonging to the same category (e.g., zip,zipped and zipper) [32,21]; (2) basic level categorization only (object recognition) and lack of fine-grained categorization [50,4,49,26,25,44,41,42,23,16,48,53,11]; (3) lack of fashion taxonomies with the needs of real-world applications for the fashion industry, possibly due to the research gap in fashion design and computer vision; (4) diverse taxonomy structures from different sources in fashion domain. To facilitate research in the areas of fashion and computer vision, our proposed ontology is built and verified by fashion experts based on their own design experiences and informed by the following four sources: (1) world-leading e-commerce fashion websites (e.g., ZARA, H&M, Gap, Uniqlo, Forever21); (2) luxury fashion brands (e.g., Prada, Chanel, Gucci); (3) trend forecasting companies (e.g., WGSN); (4) academic resources [8,3].

3 Dataset Specification and Collection

3.1 Ontology specification

We propose a unified fashion ontology (Figure 1(f)), a structured vocabulary that utilizes the basic level categories and fine-grained attributes [38]. The Fashionpedia ontology relies on similar definitions of object and attributes as previous well-known image datasets. For example, a Fashionpedia object is similar to "item" in Wikidata [45], or "object" in COCO [31] and Visual Genome [28]). In the context of Fashionpedia, objects represent common items in apparel (e.g.,

Fig. 2: Image examples with annotated segmentation masks (a-f) and fine-grained attributes (g-i)

jacket, shirt, dress). In this section, we break down each component of the Fashionpedia ontology and illustrate the construction process. With this ontology and our image dataset, a large-scale fashion knowledge graph can be built as an extended application of our dataset (more details can be found in the supplementary material).

Apparel categories. In the Fashionpedia dataset, all images are annotated with one or multiple main garments. Each main garment is also annotated with its garment parts. For example, general garment types such as jacket, dress, pants are considered as main garments. These garments also consist of several garment parts such as collars, sleeves, pockets, buttons, and embroideries. Main garments are divided into three main categories: outerwear, intimate and accessories. Garment parts also have different types: garment main parts (e.g., collars, sleeves), bra parts, closures (e.g., button, zipper) and decorations (e.g., embroidery, ruffle). On average, each image consists of 1 person, 3 main garments, 3 accessories, and 12 garment parts, each delineated by a tight segmentation mask (Figure 1(a-c)). Furthermore, each object is assigned to a synset ID in our Fashionpedia ontology.

Fine-grained attributes. Main garments and garment parts can be associated with apparel attributes (Figure 1(e)). For example, "button" is part of the main garment "jacket"; "jacket" can be linked with the silhouette attribute "symmetrical"; the garment part "button" could contain the attribute "metal" with a relationship of material. The Fashionpedia ontology provide attributes for 13 main outerwear garments categories, and 5 out of 19 garments parts ("sleeve", "neckline", "pocket", "lapel", and "collar"). Each image has 16.7 attributes on average (max 57 attributes). As with the main garments and garment parts, we canonicalize all attributes to our Fashionpedia ontology.

Relationships. Relationships can be formed between categories and attributes. There are three main types of relationships (Figure 1(e)): (1) outfits to main garments, main garments to garment parts: meronymy (part-of) relationship; (2) main garments to attributes or garment parts to attributes: these relationship types can be garment silhouette (e.g., peplum), collar nickname

(e.g., peter pan collars), garment length (e.g., knee-length), textile finishing (e.g., distressed), or textile-fabric patterns (e.g., paisley), etc.; (3) within garments, garment parts or attributes: there is a maximum of four levels of hyponymy (is-an-instance-of) relationships. For example, weft knit is an instance of knit fabric, and the fleece is an instance of weft knit.

3.2 Image Collection and Annotation Pipeline

Image collection. A total of 50,527 images were harvested from Flickr and free license photo websites, which included Unsplash, Burst by Shopify, Freestocks, Kaboompics, and Pexels. Two fashion experts verified the quality of the collected images manually. Specifically, the experts checked a scenes' diversity and made sure clothing items were visible in the images. Fdupes [33] is used to remove duplicated images. After filtering, 48,825 images were left and used to build our Fashionpedia dataset.

Annotation Pipeline. Expert annotation is often a time-consuming process. In order to accelerate the annotation process, we decoupled the work between crowd workers and experts. We divided the annotation process into the following two phases.

First, segmentation masks with apparel objects are annotated by 28 crowd workers, who were trained for 10 days before the annotation process (with prepared annotation tutorials of each apparel object, see supplementary material for details). We collected high-quality annotations by having the annotators follow the contours of garments in the image as closely as possible (See section 4.2 for annotation analysis). This polygon annotation process was monitored daily and verified weekly by a supervisor and by the authors.

Second, 15 fashion experts (graduate students in the apparel domain) were recruited to annotate the fine-grained attributes for the annotated segmentation masks. Annotators were given one mask and one attribute super-category ("textile pattern", "garment silhouette" for example) at a time. An additional two options, "not sure" and "not on the list" were added during the annotation. The option "not on the list" indicates that the expert found an attribute that is not on the proposed ontology. If "not sure" is selected, it means the expert can not identify the attribute of one mask. Common reasons for this selection include occlusion of the masks and viewing angles of the image; (for example, a top underneath a closed jacket). More details can be found in Figure 2. Each attribute supercategory is assigned to one or two fashion experts, depending on the number of masks. The annotations are also checked by another expert annotator before delivery.

We split the data into training, validation and test sets, with 45,623, 1158, 2044 images respectively. More details of the dataset creation can be found in the supplementary material.

4 Dataset Analysis

This section will discuss a detailed analysis of our dataset using the training images. We begin by discussing general image statistics, followed by an analysis of segmentation masks, categories, and attributes. We compare Fashionpedia with four other segmentation datasets, including two recent fashion datasets DeepFashion2 [11] and ModaNet [53], and two general domain datasets COCO [31] and LVIS [15].

4.1 Image Analysis

We choose to use images with high resolutions during the curating process since Fashionpedia ontology includes diverse fine-grained attributes for both garments and garment parts. The Fashionpedia training images have an average dimension of 1710 (width) $\times$ 2151 (height). Images with high resolutions are able to show apparel objects in detail, leading to more accurate and faster annotations for both segmentation masks and attributes. These high resolution images can also benefit downstream tasks such as detection and image generation. Examples of detailed annotations can be found in Figure 2.

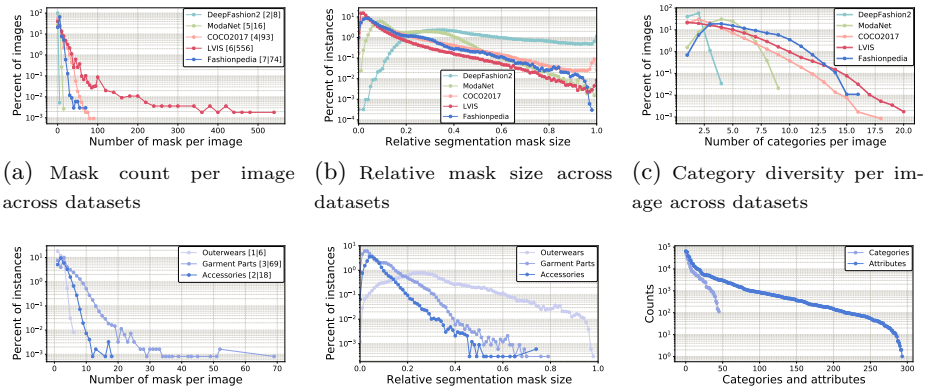
4.2 Mask Analysis

We define "masks" as one apparel instance that may have more than one separate components (the jacket in Figure 1 for example), "polygon" as a disjoint area.

Mask quantity. On average, there are 7.3 (median 7, max 74) number of masks per image in the Fashionpedia training set. Figure 3(a) shows that the Fashionpedia has the largest median value in the 5 datasets used for comparison. Fashionpedia also has the widest range among three fashion datasets, and a comparable range with COCO, which is a dataset in a general domain. Compared to ModaNet and Deepfashion2 datasets, Fashionpedia has the widest range of mask count distribution. However, COCO and LVIS maintain a wider distribution over Fashionpedia owing to their more common objects in their dataset. Figure 3(d) illustrates the distribution within the Fashionpedia dataset. One image usually contains more garment parts and accessories than outerwears.

Mask sizes. Figure 3(b) and 3(e) compares relative mask sizes within Fashionpedia and against other datasets. Ours has a similar distribution as COCO and LVIS, except for a lack of larger masks (area > 0.95). DeepFashion2 has a heavier tail, meaning it contains a larger portion of garments with a zoomed-in view. Unlike DeepFashion2, our images mainly focused on the whole ensemble of clothing. Since ModaNet focuses on outwears and accessories, it has more masks with relative area between 0.2 and 0.4. whereas ours has an additional 19 apparel parts categories. As illustrated in Figure 3(e), garment parts and accessories are relatively small compared to the outerwear (e.g., "dress", "coat").

Mask quality. Apparel categories also tend to have complex silhouettes. Table 2 shows that the Fashionpedia masks have the most complex boundaries



(a) Mask count per image across datasets

(b) Relative mask size across datasets

(c) Category diversity per image across datasets

(d) Mask count per image in Fashionpedia

(e) Relative mask size in Fashionpedia

(f) Categories and attributes distribution in Fashionpedia

Fig. 3: **Dataset statistics:** First row presents comparison among datasets. Second row presents comparison within Fashionpedia. Y-axes are in log scale. Relative segmentation mask size were calculated same as [15]. Relative mask size was rounded to precision of 2. For mask count per image comparisons (Figure 3(a) and 3(d)), legends follow [*median* | *max*] format. Values in X-axis for Figure 3(a) was discretized for better visual effect

amongst the five datasets (according to the measurement used in [15]). This suggests that our masks are better able to represent complex silhouettes of apparel categories more accurately than ModaNet and DeepFashion2. We also report the number of vertices per polygons, which is a measurement representing how granularity of the masks produced. Table 2 shows that we have the second-highest average number of vertices among five datasets, next to LVIS.

4.3 Category and Attributes Analysis

There are 46 apparel categories and 294 attributes presented in the Fashionpedia dataset. On average, each image was annotated with 7.3 instances, 5.4 categories, and 16.7 attributes. Of all the masks with categories and attributes, each mask has 3.7 attributes on average (max 14 attributes). Fashionpedia has the most diverse number of categories within one image among three fashion datasets, while comparable to COCO (Figure 3(c)), since we provide a comprehensive ontology for the annotation. In addition, Figure 3(f) shows the distributions of categories and attributes in the training set, and highlights the long-tailed nature of our data.

During the fine-grained attributes annotation process, we also ask the experts to choose "not sure" if they are uncertain to make a decision, "not on the list" if they find another attribute that not provided. the majority of "not sure" comes from three attributes superclasses, namely "Opening Type", "Waistline",

Table 2: Comparison of segmentation mask complexity amongst segmentation datasets in both fashion and general domain (COCO 2017 instance training data was used). Each statistic (mean and median) represents a bootstrapped 95% confidence interval following [15]. Boundary complexity was calculated according to [15,2]. Reported mask boundary complexity for COCO and LVIS was different compared with [15] due to different image resolution and image sets. The number of vertices per polygon is calculated as the number of vertices in one polygon. Polygon is defined as one disjoint area. Masks (and polygons) with zero area were ignored

Dataset	**Boundary complexity**		**No. of vertices per polygon**		**Images count**
	mean	median	mean	median	
COCO [31]	6.65 - 6.66	6.07 - 6.08	21.14 - 21.21	15.96 - 16.04	118,287
LVIS [15]	6.78 - 6.80	5.89 - 5.91	**35.77 - 35.95**	**22.91 - 23.09**	57,263
ModaNet [53]	5.87 - 5.89	5.26 - 5.27	22.50 - 22.60	18.95 - 19.05	52,377
Deepfashion2 [11]	4.63 - 4.64	4.45 - 4.46	14.68 - 14.75	8.96 - 9.04	**191,960**
Fashionpedia	**8.36 - 8.39**	**7.35 - 7.37**	31.82 - 32.01	20.90 - 21.10	45, 623

"Length". Since there are masks only show a limited portion of apparel (a top inside a jacket for example), the annotators are not sure how to identify those attributes due to occlusion or viewpoint discrepancies. Less than 15% of masks for each attribute superclasses account for "not on the list", which illustrates the comprehensiveness of our proposed ontology (see supplementary material for more details of the extra dataset analysis).

5 Evaluation Protocol and Baselines

5.1 Evaluation metric

In object detection, a true positive (TP) for each category c is defined as a single detected object that matches a ground truth object with a Intersection over Union (IoU) over a threshold τ_{IoU}. COCO's main evaluation metric uses average precision averaged across all 10 IoU thresholds $\tau_{\text{IoU}} \in [0.5 : 0.05 : 0.95]$ and all 80 categories. We denote such metric as AP_{IoU}.

In the case of instance segmentation and attribute localization , we extend standard COCO metric by adding one more constraint: the macro F_1 score for predicted attributes of single detected object with category c (see supplementary material for the average choice of f1-score). We denote the F_1 threshold as τ_{F_1}, and it has the same range as τ_{IoU} ($\tau_{\text{F}_1} \in [0.5 : 0.05 : 0.95]$). The main metric $\text{AP}_{\text{IoU+F}_1}$ reports averaged precision score across all 10 IoU thresholds, all 10 macro F_1 scores, and all the categories. Our evaluation API, code and trained models are available at: https://fashionpedia.github.io/home/Model_and_API.html.

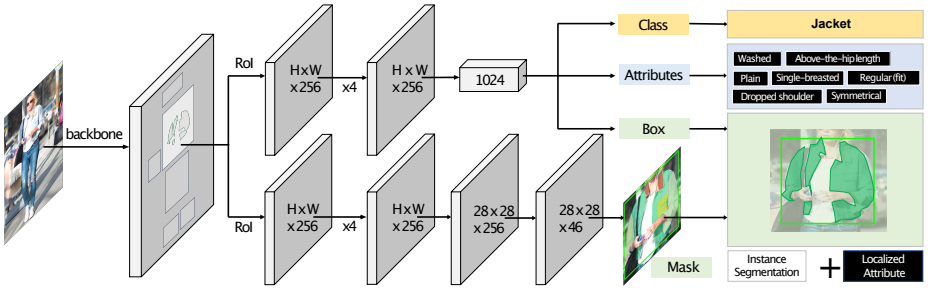


Fig. 4: Attribute-Mask R-CNN adds a multi-label attribute prediction head upon Mask R-CNN for instance segmentation with attribute localization

5.2 Attribute-Mask R-CNN

We perform two tasks on Fashionpedia: (1) apparel instance segmentation (ignoring attributes); (2) instance segmentation with attribute localization. To better facilitate research on Fashionpedia, we design a strong baseline model named Attribute-Mask R-CNN that is built upon Mask R-CNN [17]. As illustrated in Figure 4, we extend Mask R-CNN heads to include an additional multi-label attribute prediction head, which is trained with sigmoid cross-entropy loss. Attribute-Mask R-CNN can be trained end-to-end for jointly performing instance segmentation and localized attribute recognition.

We leverage different backbones including ResNet-50/101 (R50/101) [18] with feature pyramid network (FPN) [30] and SpineNet-49/96/143 [5]. The input image is resized to 1024 of the longer edge to feed all networks except SpineNet-143. For SpineNet-143, we instead use the input size of 1280. During training, other than standard random horizontal flipping and cropping, we use large scale jittering that resizes an image to a random ratio between $(0.5, 2.0)$ of the target input image size, following Zoph *et al*. [54]. We use an open-sourced Tensorflow [1] code base[1] for implementation and all models are trained with a batch size of 256. We follow the standard training schedule of $1\times$ (5625 iterations), $2\times$ (11250 iterations), $3\times$ (16875 iterations) and $6\times$ (33750 iterations) in Detectron2 [47], with linear learning rate scaling suggested by Goyal *et al*. [13].

5.3 Results Discussion

Attribute-Mask R-CNN. From results in Table 3, we have the following observations: (1) Our baseline models achieve promising performance on challenging Fashionpedia dataset. (2) There is a significant drop (*e.g.*, from 48.7 to 35.7 for SpineNet-143) in box AP if we add τ_{F_1} as another constraint for true positive. This is further verified by per super-category mask results in Table 4. This suggests that joint instance segmentation and attribute localization is a significantly more difficult task than instance segmentation, leaving much more room for future improvements.

Main apparel detection analysis. We also provide in-depth detector analysis following COCO detection challenge evaluation [31] inspired by Hoiem *et al*. [20]. Figure 5 illustrates a detailed breakdown of bounding boxes false positives produced by the detectors.

[1] https://github.com/tensorflow/tpu/tree/master/models/official/detection

Table 3: Baseline results of Mask R-CNN and Attribute-Mask R-CNN on Fashionpedia. The big performance gap between AP_{IoU} and AP_{IoU+F_1} suggests the challenging nature of our proposed task

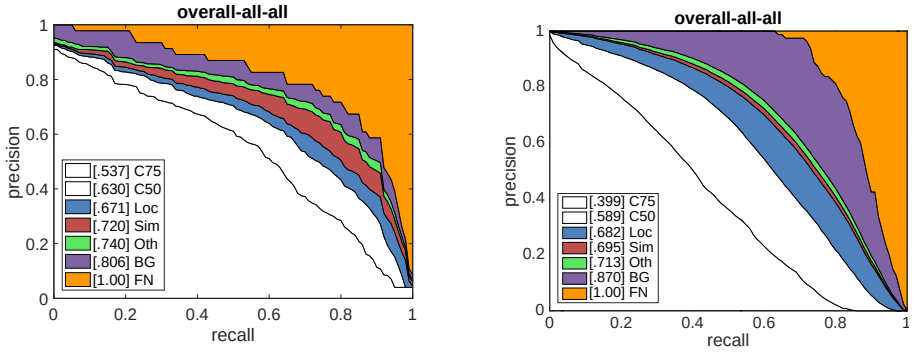
Backbone	Schedule	FLOPs(B)	params(M)	$AP^{box}_{IoU/IoU+F_1}$	$AP^{mask}_{IoU/IoU+F_1}$
R50-FPN	1×	296.7	46.4	38.7 / 26.6	34.3 / 25.5
	2×			41.6 / 29.3	38.1 / 28.5
	3×			43.4 / 30.7	39.2 / 29.5
	6×			42.9 / 31.2	38.9 / 30.2
R101-FPN	1×	374.3	65.4	41.0 / 28.6	36.7 / 27.6
	2×			43.5 / 31.0	39.2 / 29.8
	3×			44.9 / 32.8	40.7 / 31.4
	6×			44.3 / 32.9	39.7 / 31.3
SpineNet-49	**6**×	267.2	40.8	43.7 / 32.4	39.6 / 31.4
SpineNet-96		314.0	55.2	46.4 / 34.0	41.2 / 31.8
SpineNet-143		**498.0**	**79.2**	**48.7 / 35.7**	**43.1 / 33.3**

Table 4: Per super-category results (for masks) using Attribute-Mask R-CNN with SpineNet-143 backbone. We follow the same COCO sub-metrics for overall and three super-categories for apparel objects. Result format follows [AP_{IoU} / AP_{IoU+F_1}] or [AR_{IoU} / AR_{IoU+F_1}] (see supplementary material for per-class results)

Category	AP	AP50	AP75	APl	APm	APs
overall	43.1 / 33.3	60.3 / 42.3	47.6 / 37.6	50.0 / 35.4	40.2 / 27.0	17.3 / 9.4
outerwear	64.1 / 40.7	77.4 / 49.0	72.9 / 46.2	67.1 / 43.0	44.4 / 29.3	19.0 / 4.4
parts	19.3 / 13.4	35.5 / 20.8	18.4 / 14.4	28.3 / 14.5	23.9 / 16.4	12.5 / 9.8
accessory	56.1 / -	77.9 / -	63.9 / -	57.5 / -	60.5 / -	25.0 / -

Figure 5(a) and 5(b) compare two detectors trained on Fashionpedia and COCO. Errors of the COCO detector are dominated by imperfect localization (AP is increased by 28.3 from overall AP at $\tau_{IoU} = 0.75$) and background confusion (+15.7) (5(b)). Unlike the COCO detector, no mistake in particular dominates the errors produced by the Fashionpedia detector. Figure 5(a) shows that there are errors from localization (+13.4.0), classification (+6.9), background confusions (+6.6). Due to the space constraint, we leave super-category analysis in the supplementary material.

Prediction visualization. Baseline outputs (with both segmentation masks and localized attributes) are also visualized in Figure 6. Our Attribute-Mask R-CNN achieves good results even for small objects like shoes and glasses. Model can correctly predict fine-grained attributes for some masks on one hand (e.g., Figure 6 second image at top row Sleeve 1). On the other hand, it also incorrectly predict the wrong nickname (welt) to pocket (e.g., Figure 6 third image at top row Pocket 1). These results show that there is headroom remaining for future development of more advanced computer vision models on this task (see supplementary material for more details of the baseline analysis).



(a) Fashionpedia (b) COCO (2015 challenge winner)

Fig. 5: Main apparel detectors analysis. Each plot shows 7 precision recall curves where each evaluation setting is more permissive than the previous. Specifically, **C75**: strict IoU ($\tau_{\text{IoU}} = 0.75$); **C50**: PASCAL IoU ($\tau_{\text{IoU}} = 0.5$); **Loc**: localization errors ignored ($\tau_{\text{IoU}} = 0.1$); **Sim**: supercategory False Positives (FPs) removed; **Oth**: category FPs removed; **BG**: background (and class confusion) FPs removed; **FN**: False Negatives are removed. Two plots are a comparison between two detectors trained on Fashionpedia and COCO respectively. The results are averaged over all categories. Legends present the area under each curve (corresponds to AP metric) in brackets as well

6 Conclusion

In this work, we focus on a new task that unifies instance segmentation and attribute recognition. To solve challenging problems entailed in this task, we introduced the Fashionpedia ontology and dataset. To the best of our knowledge, Fashionpedia is the first dataset that combines part-level segmentation masks with fine-grained attributes. We presented Attribute-Mask R-CNN, a novel model for this task, along with a novel evaluation metric. We expect models trained on Fashionpedia can be applied to many applications including better product recommendation in online shopping, enhanced visual search results, and resolving ambiguous fashion-related words for text queries. We hope Fashionpedia will contribute to the advances in fine-grained image understanding in the fashion domain.

Acknowledgements

This research was partially supported by a Google Faculty Research Award. We thank Kavita Bala, Carla Gomes, Dustin Hwang, Rohun Tripathi, Omid Poursaeed, Hector Liu, and Nayanathara Palanivel, Konstantin Lopuhin for their helpful feedback and discussion in the development of Fashionpedia dataset. We also thank Zeqi Gu, Fisher Yu, Wenqi Xian, Chao Suo, Junwen Bai, Paul Upchurch, Anmol Kabra, and Brendan Rappazzo for their help developing the fine-grained attribute annotation tool.

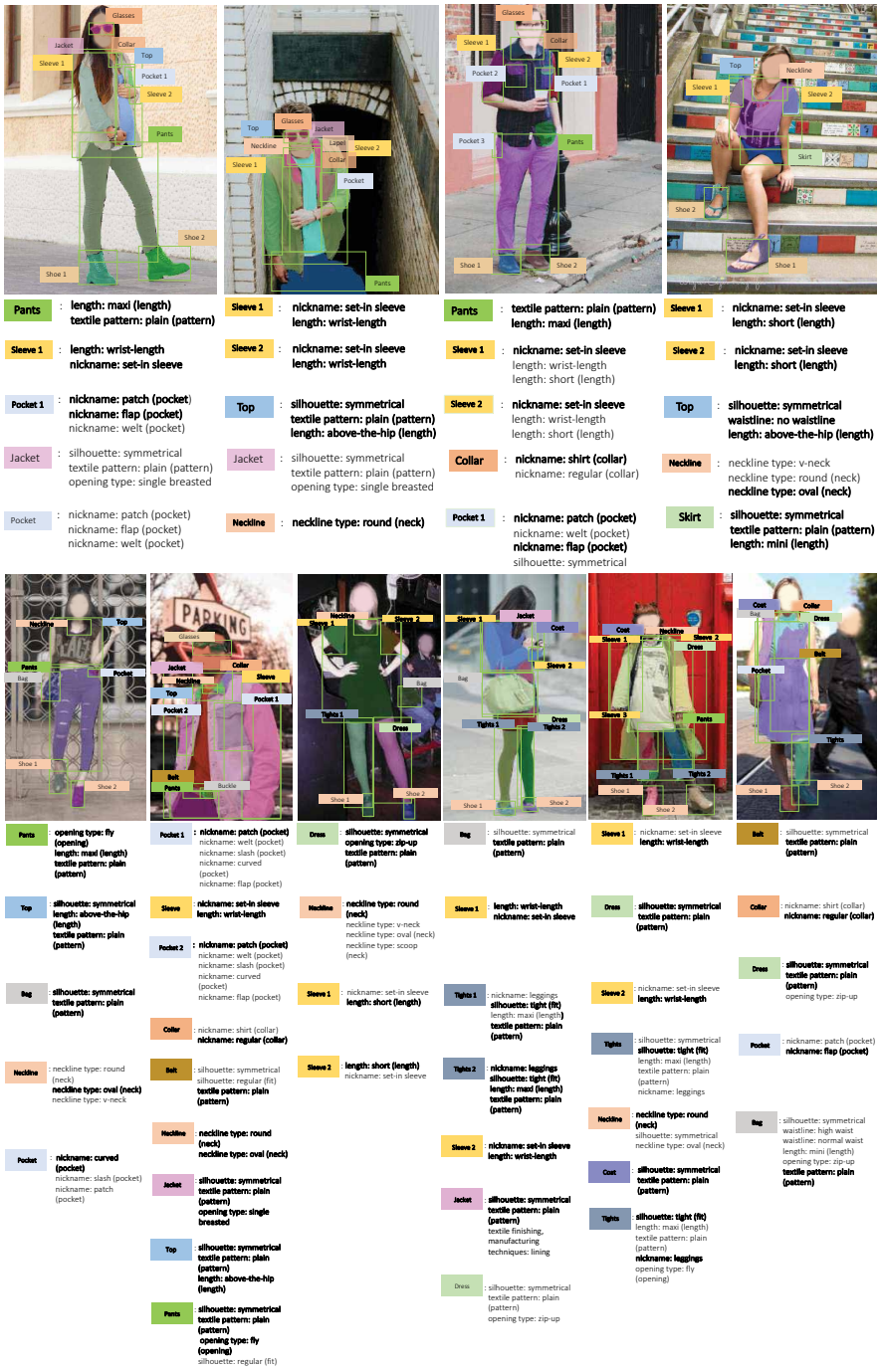


Fig. 6: Attribute-Mask R-CNN results on the Fashionpedia validation set. Masks, bounding boxes, and apparel categories (category score > 0.6) are shown. The localized attributes from the top 5 masks (that contain attributes) on each image are also shown. Correctly predicted categories and localized attributes are bolded

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: OSDI (2016) 11
2. Attneave, F., Arnoult, M.D.: The quantitative study of shape and pattern perception. Psychological bulletin (1956) 10
3. Bloomsbury.com: Fashion photography archive, retrieved May 9, 2019 from https://www.bloomsbury.com/dr/digital-resources/products/fashion-photography-archive/ 5
4. Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., Van Gool, L.: Apparel classification with style. In: ACCV (2012) 4, 5
5. 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 (2020) 11
6. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009) 1
7. FashionAI: Retrieved May 9, 2019 from http://fashionai.alibaba.com/ 4
8. Fashionary.org: Fashionpedia - the visual dictionary of fashion design, retrieved May 9, 2019 from https://fashionary.org/products/fashionpedia 5
9. Ferrari, V., Zisserman, A.: Learning visual attributes. In: Advances in neural information processing systems (2008) 1, 2
10. Fu, C.Y., Berg, T.L., Berg, A.C.: Imp: Instance mask projection for high accuracy semantic segmentation of things. arXiv preprint arXiv:1906.06597 (2019) 5
11. Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P.: Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: CVPR (2019) 2, 4, 5, 8, 10
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014) 1
13. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017) 11
14. Guo, S., Huang, W., Zhang, X., Srikhanta, P., Cui, Y., Li, Y., Adam, H., Scott, M.R., Belongie, S.: The imaterialist fashion attribute dataset. In: ICCV Workshops (2019) 1, 4, 5
15. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: CVPR (2019) 3, 8, 9, 10
16. Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S.: Automatic spatially-aware fashion concept discovery. In: ICCV (2017) 4, 5
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017) 1, 11
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 11
19. He, R., McAuley, J.: Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In: WWW (2016) 4
20. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: ECCV (2012) 11
21. Hsiao, W.L., Grauman, K.: Learning the latent look: Unsupervised discovery of a style-coherent embedding from fashion images. In: ICCV (2017) 4, 5
22. Huang, J., Feris, R., Chen, Q., Yan, S.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: ICCV (2015) 4, 5

23. Inoue, N., Simo-Serra, E., Yamasaki, T., Ishikawa, H.: Multi-label fashion image classification with minimal human supervision. In: ICCV (2017) 4, 5
24. Kendall, E.F., McGuinness, D.L.: Ontology engineering. Synthesis Lectures on The Semantic Web: Theory and Technology (2019) 3
25. Kiapour, M.H., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: Matching street clothing photos in online shops. In: ICCV (2015) 4, 5
26. Kiapour, M.H., Yamaguchi, K., Berg, A.C., Berg, T.L.: Hipster wars: Discovering elements of fashion styles. In: ECCV (2014) 4, 5
27. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: CVPR (2019) 3
28. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. IJCV (2017) 3, 4, 5
29. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV (2009) 1
30. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017) 11
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 3, 5, 8, 10, 11
32. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR (2016) 1, 4, 5
33. Lopez, A.: Fdupes is a program for identifying or deleting duplicate files residing within specified directories., retrieved May 9, 2019 from https://github.com/adrianlopezroche/fdupes 7
34. Mall, U., Matzen, K., Hariharan, B., Snavely, N., Bala, K.: Geostyle: Discovering fashion trends and events. In: ICCV (2019) 5
35. Matzen, K., Bala, K., Snavely, N.: StreetStyle: Exploring world-wide clothing styles from millions of photos. arXiv preprint arXiv:1706.01869 (2017) 4
36. Parikh, D., Grauman, K.: Relative attributes. In: ICCV (2011) 1
37. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems (2015) 1
38. Rosch, E.: Cognitive representations of semantic categories. Journal of experimental psychology: General (1975) 2, 5
39. Rubio, A., Yu, L., Simo-Serra, E., Moreno-Noguer, F.: Multi-modal embedding for main product detection in fashion. In: ICCV (2017) 4, 5
40. Simo-Serra, E., Fidler, S., Moreno-Noguer, F., Urtasun, R.: Neuroaesthetics in fashion: Modeling the perception of fashionability. In: CVPR (2015) 4
41. Simo-Serra, E., Ishikawa, H.: Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In: CVPR (2016) 4, 5
42. Takagi, M., Simo-Serra, E., Iizuka, S., Ishikawa, H.: What makes a style: Experimental analysis of fashion prediction. In: ICCV (2017) 4, 5
43. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: CVPR (2018) 2, 4
44. Vittayakorn, S., Yamaguchi, K., Berg, A.C., Berg, T.L.: Runway to realway: Visual analysis of fashion. In: WACV (2015) 4, 5
45. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledge base (2014) 5

46. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011) 2, 4
47. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019) 11
48. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017) 4, 5
49. Yamaguchi, K., Berg, T.L., Ortiz, L.E.: Chic or social: Visual popularity analysis in online fashion networks. In: ACM MM (2014) 4, 5
50. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L.: Parsing clothing in fashion photographs. In: CVPR (2012) 4, 5
51. Yu, A., Grauman, K.: Semantic jitter: Dense supervision for visual comparisons via synthetic images. In: ICCV (2017) 4
52. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: CVPR (2020) 1
53. Zheng, S., Yang, F., Kiapour, M.H., Piramuthu, R.: Modanet: A large-scale street fashion dataset with polygon annotations. In: ACM MM (2018) 2, 4, 5, 8, 10
54. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.V.: Rethinking pre-training and self-training. arXiv preprint arXiv:2006.06882 (2020) 11