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

1 Supplementary Material

In our work we presented the new task of instance segmentation with attribute localization. We introduced a new ontology and dataset, Fashionpedia, to further describe the various aspects of this task. We also proposed a novel evaluation metric and Attribute-Mask R-CNN model for this task. In the supplemental material, we provide the following items that shed further insight on these contributions:

- More comprehensive experimental results (§ 1.1)
- An extended discussion of Fashionpedia ontology and potential knowledge graph applications (§ 1.2)
- More details of dataset analysis (§ 1.3)
- Additional information of annotation process (§ 1.4)
- Discussion (§ 1.5)

1.1 Attribute-Mask R-CNN

Per-class evaluation. Fig. 1 presents detailed evaluation results per supercategory and per category. In Fig. 1, we follow the same metrics from COCO leaderboard (AP, AP50, AP75, APl, APm, APs, AR@1, AR@10, AR@100, ARs@100, ARm@100, ARl@100), with τ_{IoU} and τ_{F1} if possible. Fig. 1 shows that metrics considering both constraint τ_{IoU} and τ_{F1} are always lower than using τ_{IoU} alone across all the supercategories and categories. This further demonstrates the challenging aspect of our proposed task.

In general, categories belong to "garment parts" have a lower AP and AR, comparing with "outerwear" and "accessories".

A detailed breakdown of detection errors is presented in Fig. 2 for supercategories and three main categories. In terms of supercategories in Fashionpedia, "outerwear" errors are dominated by within supercategory class confusions (Fig. 2(a)). Within this supercategory class, ignoring localization errors would only raise AP slightly from 77.5 to 79.1 (+1.6). A similar trend can be observed in class "skirt", which belongs to "outerwear" (Fig. 2(d)). Detection errors of "part" (Fig. 2(b) 2(e)) and "accessory"(Fig. 2(c) 2(f)) on the other hand, are

* equal contribution.

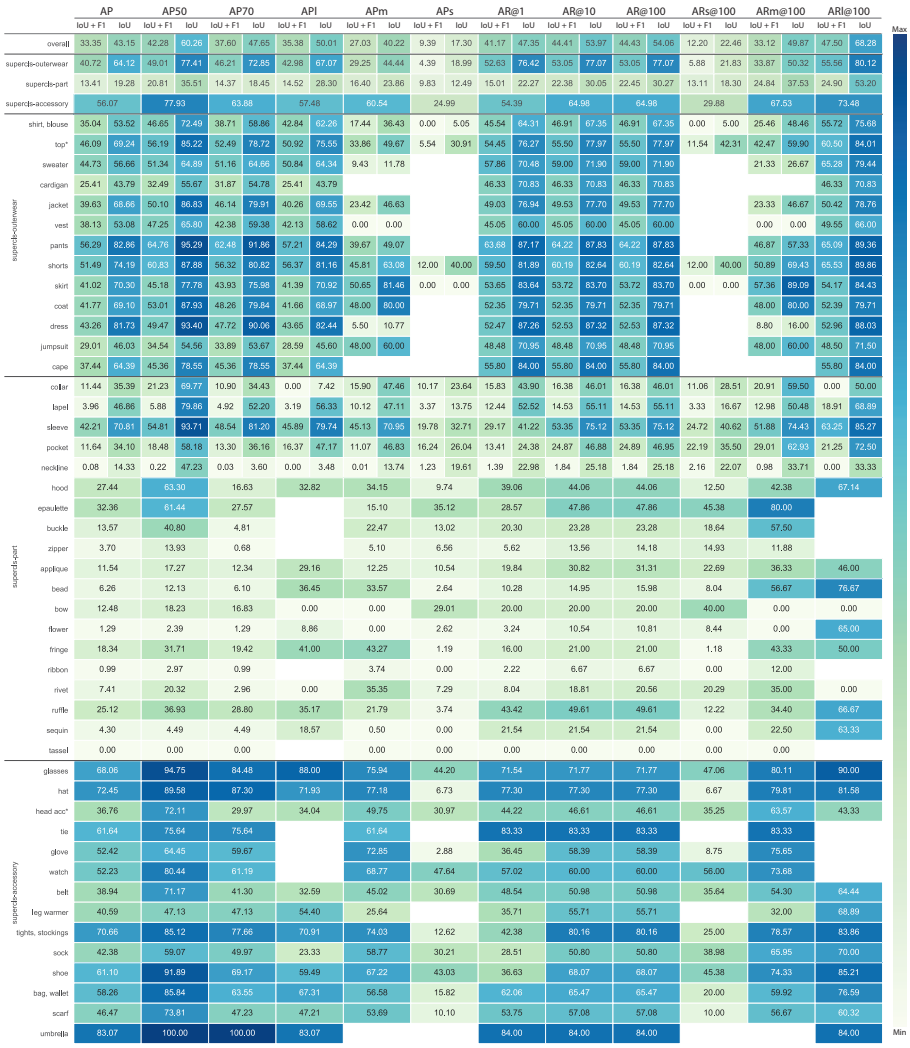
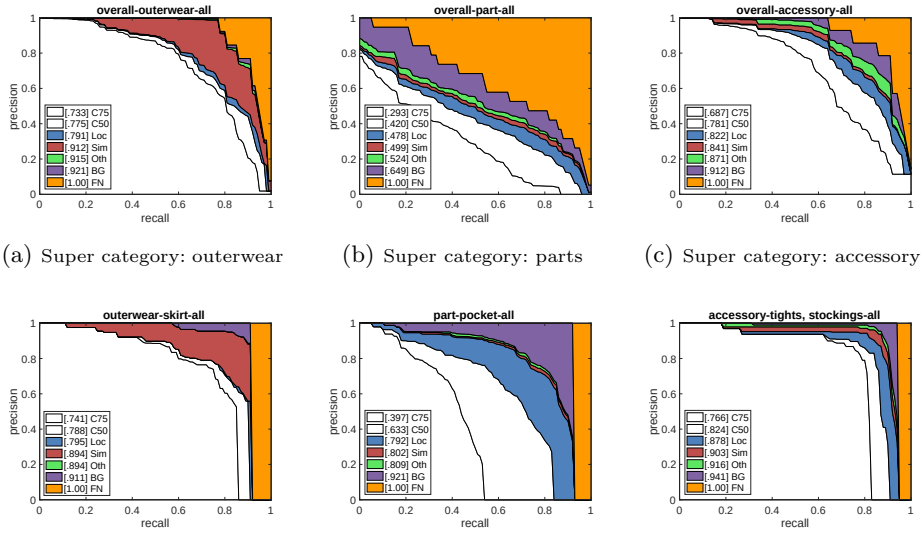


Fig. 1: Detailed results (for masks) using Mask R-CNN with SpineNet-143 backbone. We present the same metrics as COCO leaderboard for overall categories, three super categories for apparel objects, and 46 fine-grained apparel categories. We use both constraints (for example, AP_{IoU} and AP_{IoU+F_1}) if possible. For categories without attributes, the value represents AP_{IoU} or AR_{IoU}. "top" is short for "top, t-shirt, sweatshirt". "head acc" is short for "headband, head covering, hair accessory"

dominated by both background confusion and localization. "part" also has a lower AP in general, compared with other two super-categories. A possible reason is that objects belong to "part" usually have smaller sizes and lower counts.



(a) Super category: outerwear (b) Super category: parts (c) Super category: accessory

(d) Outerwear: skirt (e) Parts: pocket (f) Accessory: tights, stockings

Fig. 2: 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. The first row *(overall-[supercategory]-[size])* contains results for three supercategories in Fashionpedia; the second row *([supercategory]-[category]-[size])* shows results for three fine-grained categories (one per supercategory). Legends present the area under each curve (corresponds to AP metric) in brackets as well

F1 score calculation. Since we measure the f1 score of predicted attributes and groundtruth attributes per mask, we consider the both options of multi-label multi-class classification with 294 classes for one instance, and binary classification for 294 instances. Multi-label multi-class classification is a straightforward task, as it is a common setting for most of the fine-grained classification tasks. In binary classification scenario, we consider the 1 and 0 of the multi-hot encoding of both results and ground-truth labels as the positive and negative classes respectively. There are also two averaging choices: "micro" and "macro". "Micro"

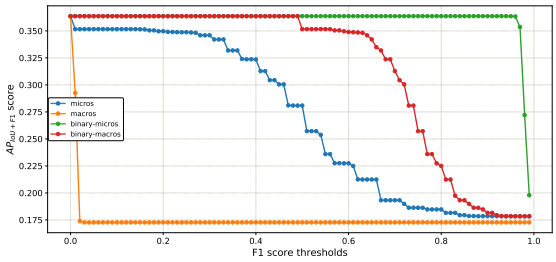


Fig. 3: $AP^{F1=\tau_{F1}}_{IoU+F1}$ score with different τ_{F1}. The value presented are average over $\tau_{IoU} \in [0.5 : 0.05 : 0.95]$. We use "binary-macro" as our main metric

averaging calculates the score globally by counting the total true positives, false negatives and false positives. "Macro" averaging calculates the metrics for each attribute class and reports the unweighted mean. In sum, there are four options of f1-score averaging methods: 1) "micro", 2) "macro", 3) "binary-micro", 4) "binary-macro".

As shown in Fig. 3, we present the $AP^{F1=\tau_{F1}}_{IoU+F1}$, with τ_{IoU} averaged in the range of $[0.5 : 0.05 : 0.95]$. τ_{F1} is increased from 0.0 to 1.0 with a increment of 0.01. Fig. 3 illustrates that as the value of τ_{F1} increases, $AP^{F1=\tau_{F1}}_{IoU+F1}$ decreases in different rates given different choices of f1 score calculation. There are 294 attributes in total, and an average of 3.7 attributes per mask in Fashionpedia training data. It's not surprising to observe that"Binary-micro" produces high f1-scores in general (higher than 0.97), as the AP_{IoU+F1} score only decreases if the $\tau_{F1} \geq 0.97$. On the other hand, "macro" averaging in multi-label multi-class classification scenario gives us extremely low f1-scores $(0.01 - 0.03)$. This further demonstrates the room for improvement for localized attribute classification task. We used "binary-macros" as our main metric.

Result visualization. Figure 4 shows that our simple baseline model can detect most of the apparel categories correctly. However, it also produces false positives sometimes. For example, it segments legs as tights and stockings (Figure 4(f) Tights). A possible reason is that both objects have the same shape and stockings are worn on the legs.

Predicting fine-grained attributes, on the other hand, is a more challenging problem for the baseline model. We summarize several issues: (1) predict more attributes than needed: (Figure 4(a) Neckline, (b) Collar, (c) Sleeve); (2) fail to distinguish among fine-grained attributes: for example, dropped-shoulder sleeve (ground truth) v.s. set-in sleeve (predicted) (Figure 4(e) Sleeve 1); (3) other false positives: Figure 4(e) Dress has a double-breasted opening, yet the model predicted it as the zip opening.

These results further show that there are rooms for improvement and future development of more advanced computer vision models on this instance segmentation with attribute localization task.

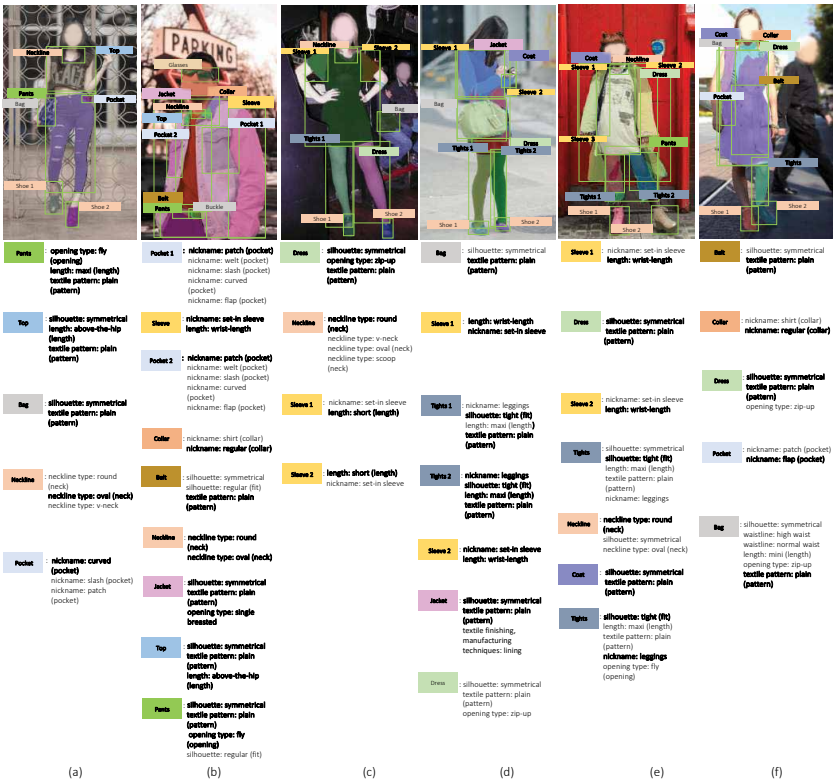


Fig. 4: Baseline results on the Fashionpedia validation set. Masks, bounding boxes, and apparel categories (category score > 0.6) are shown. Attributes from top 10 masks (that contain attributes) from each image are also shown. Correct predictions of objects and attributes are bolded

Result visualization on other datasets. Other fashion datasets such as ModaNet and DeepFashion2 also contain instance segmentation masks. Aside from the results presented in the main paper (see Table 3 of the main paper) on Fashionpedia, we present the qualitative analysis on the segmentation masks generated among Fashionpedia(Fig. 4), ModaNet(Fig. 5(a-f)) and DeepFashion2(Fig. 5(g-l)) datasets. Photos of the first row in Fig. 5 are from ModaNet. They show that the quality of the generated masks on ModaNet is fairly good and comparable to Fashionpedia in general (Fig. 5(a)). We also have a couple of observations of the failure cases: (1) fail to detect apparel objects: for example, the shoe from Fig. 5(c) is not detected. Parts of the pants (Fig. 5(c)) and coat (Fig. 5(d)) are not detected; (2) fail to detect some categories: Fig. 5(e) shows that the shoes on the shoe rack and right foot are not detected, possibly due to a lack of such instances in the ModaNet training dataset. Similar to

Fig. 5: Baseline results on ModaNet and DeepFashion2 validation set

Fig. 6: Generated masks on online-shopping images [2]. (a) and (b) show the same types of shoes in different settings. Our model correctly detects and categorizes the pair of shoes worn by a fashion model, yet mistakenly detects shoes as jacket and a bag in (b)

Fashionpedia, ModaNet mostly consist of street style images. See Fig. 6(b) for example predictions from model trained on Fashionpedia; 3) close-up images: ModaNet contains mostly full-body images. This might be the possible reason to the decreased quality of predicted masks on close-up shot like Fig. 5(f).

For DeepFashion 2 (Fig. 5(g,h,k)), the generated segmentation masks tends to not tightly follow the contours of garments in the images. The main reason possibly is that the average number of vertices per polygon is 14.7 for Deepfashion2, which is lower than Fashionpedia and ModaNet (see Table 2 in the main text). Our qualitative analysis also shows that: 1) the model will generate the segmentation masks of pants (Fig. 5(i)) and tops (Fig. 5(j)) that are not visible in the images. Both of them are covered by a jacket. And we find that in DeepFashion 2, some part of the garments which is covered by other objects are indeed annotated with segmentation masks; 2) better performance on objects that are not on human body (Fig. 5(l)): DeepFashion 2 contains many commercial-customer image pairs (both images with and without human body) in the training dataset. In contrast, both Fashionpedia and ModaNet contain more images with human body than images without human body in the training datasets.

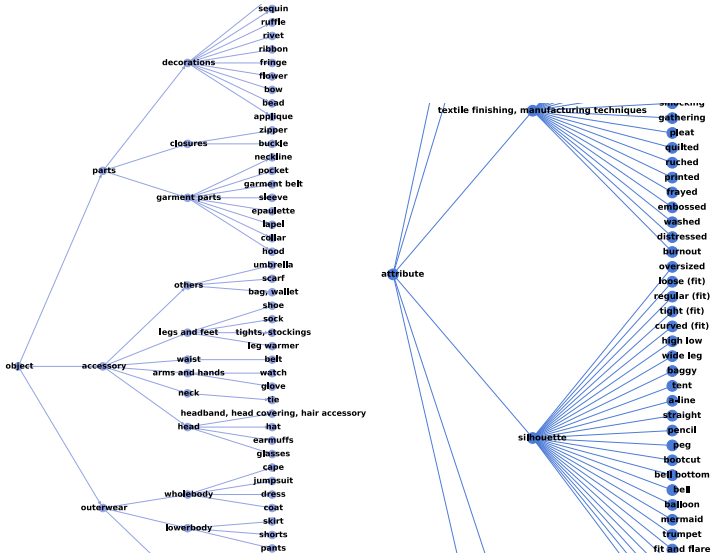
Generalization to the other image domains. For Fashionpedia, we also inference on images found in online shopping websites, which usually displays a single apparel category, with or without a fashion model. We found out that the learned model works reasonably well if the apparel item is worn by a model (Fig. 6).

1.2 Fashionpedia Ontology and Knowledge Graph

Fig. 7 presents our Fashionpedia ontology in detail. Fig. 8 and 9 displays the training data mask counts per category and per attributes. Utilizing the proposed ontology and the image dataset, a large-scale fashion knowledge graph can be constructed to represent the fashion world in the product level. Fig. 10 illustrates a subset of the Fashionpedia knowledge graph.

Apparel graphs. Integrating the main garments, garment parts, attributes, and relationships presented in one outfit ensemble, we can create an apparel graph representation for each outfit in an image. Each apparel graph is a structured representation of an outfit ensemble, containing certain types of garments. Nodes in the graph represent the main garments, garment parts, and attributes. Main garments and garment parts are linked to their respective attributes through different types of relationships. Figure 11 shows more image examples with apparel graphs.

Fashionpedia knowledge graph. While apparel graphs are localized representations of certain outfit ensembles in fashion images, we can also create a single Fashionpedia knowledge graph (Figure 10). The Fashionpedia knowledge graph is the union of all apparel graphs and includes entire main garments, garment parts, attributes, and relationships in the dataset. In this way, we are able to represent and understand fashion images in a more structured way.



(a) Categories (b) Attributes (partially)

Fig. 7: Apparel categories (a) and fine-grained attributes (b) hierarchy in Fashionpedia

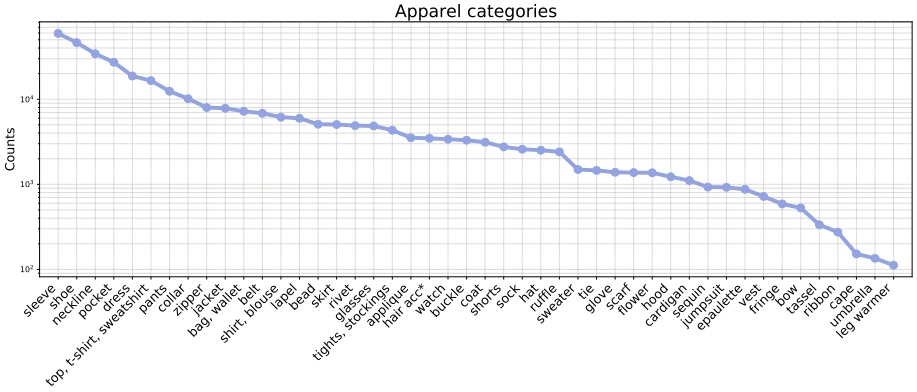


Fig. 8: Mask counts per apparel categories in training data. "head acc" is short for "headband, head covering, hair accessory"

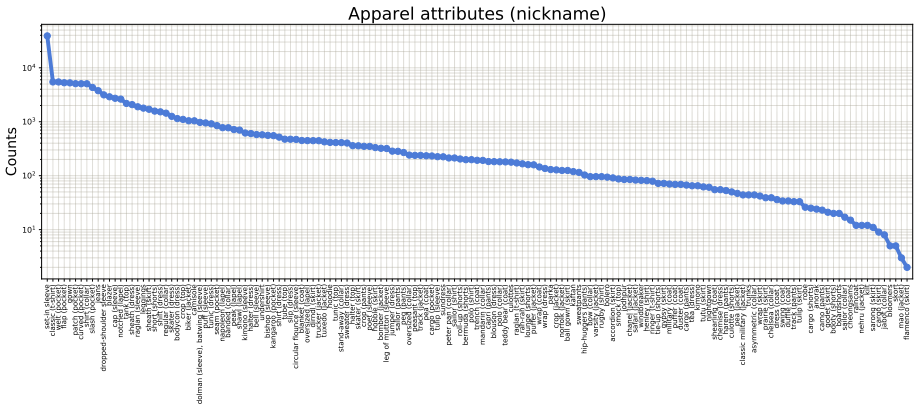
We expect our Fashionpedia knowledge graph and the database to have applicability to extending the existing knowledge graph (such as WikiData [1]) with novel domain-specific knowledge, improving the underlying fashion product recommendation system, enhancing search engine's results for fashion visual search, resolving ambiguous fashion-related words for text search, and more.

1.3 Dataset Analysis

Fig. 11 shows more annotation examples, represented in the exploded views of annotation diagrams. Table 1 displays the details about "not sure" and "not on the list" results during attribute annotation process. We present the result per super-categories of attributes. Label "not sure" means the expert annotator is uncertain about the choice given the segmentation mask. "Not on the list" means the annotator is certain that the given mask presents another attributes that is not presented in the Fashionpedia ontology. Other than "nicknames" (which is the specific name for a certain apparel category), less than 6% of the total masks account for the "not on the list" category.

Fig. 12 and 13 also compare Fashionpedia and other images datasets in terms of image size and vertices per polygons.

We compare image resolutions between Fashionpedia and four other segmentation datasets (COCO and LVIS share the same images). Fig. 12 shows that images in Fashionpedia have the most diverse image width and height. While ModaNet has the most consistent resolutions of images. Note that high resolution images will burden the data loading process of training. With that in mind, we will release our dataset in both the resized and the original versions.



(a)

Fig. 9: Mask counts per attributes in training data, grouped by super categories. Best viewed digitally

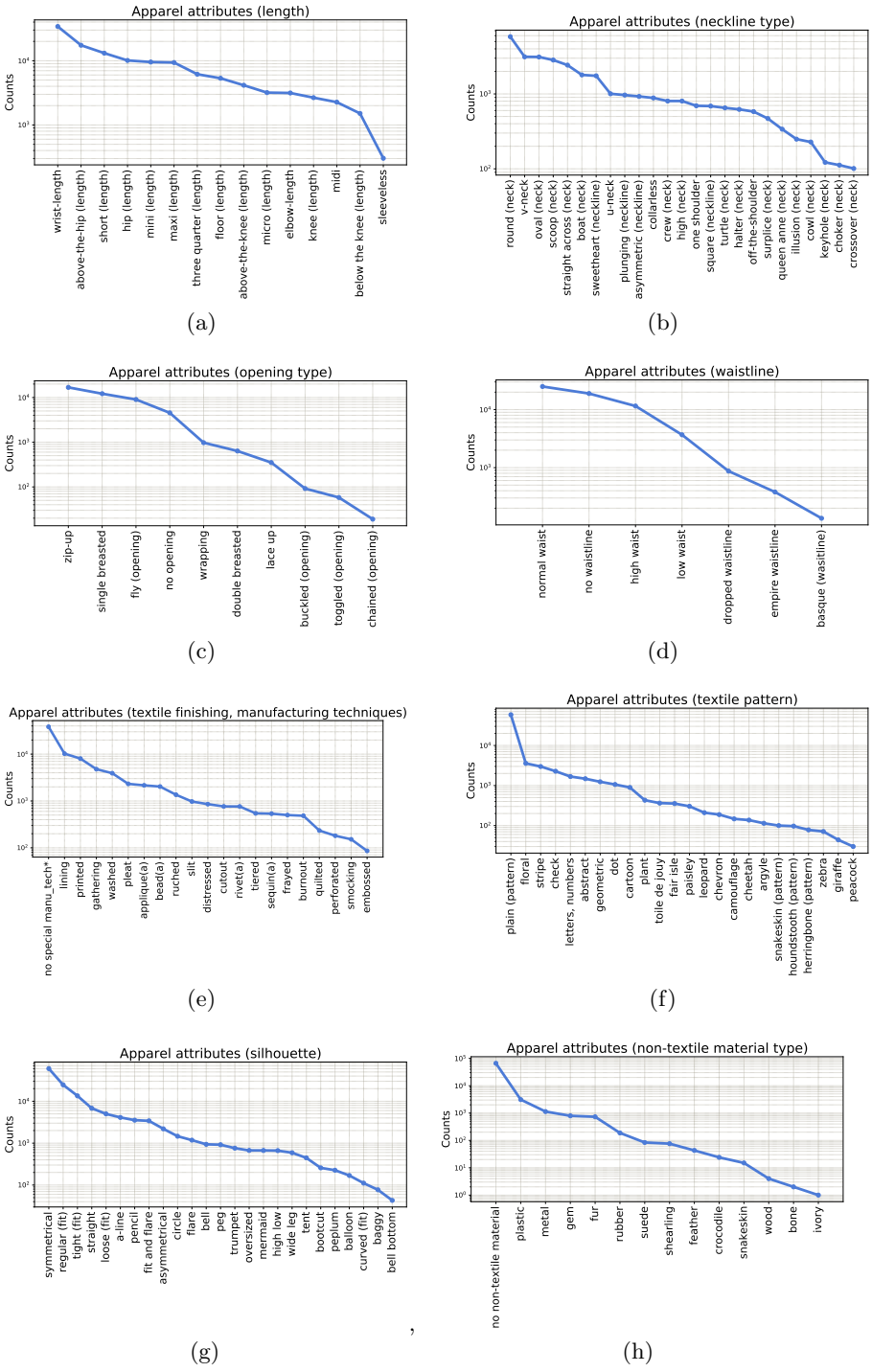


Fig. 9: Mask counts per attributes in training data, grouped by super categories (cont.). Best viewed digitally

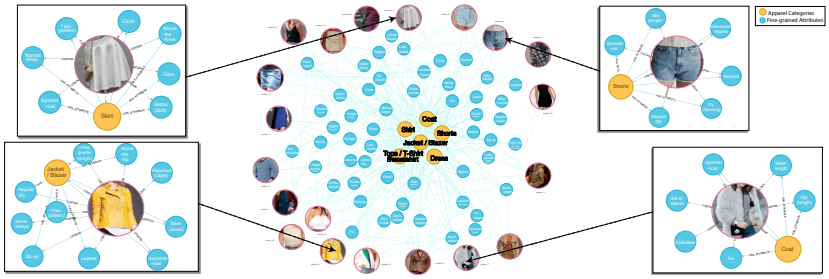


Fig. 10: Fashionpedia Knowledge Graph: we present a subset of the Fashionpedia knowledge graph by aggregating 20 annotated products. The knowledge graph can be used as a tool for generating structural information

We also report the distribution of number of vertices per polygons in Fig. 13. This measures the annotation effort in mask annotation. Fashionpedia has the second-widest range, next to LVIS.

1.4 Fashionpedia dataset creation details

Image collection. To avoid photo bias, all the images are randomly collected from Flickr and free license stock photo websites (Unsplash, Burst by Shopify, Free stocks, Kaboompics, and Pexels). The collected images are further verified manually by two fashion experts. Specifically, they check the scenes' diversity

Table 1: Percentage of attributes in Fashionpedia broken down by super-class. "Tex finish, manu-tech." is short for "Textile finishing, Manufacturing techniques". Summaries of "not sure" and "not on the list" during attributes annotations are also presented. It was calculated by the counts divided by the total masks with attributes. "not sure" is mainly due to occlusion inside the images, which cause some super-classes (such as waistline, opening type, and length) are unidentifiable in the images. The percentage of "not on the list" is less than 15%. This demonstrates the comprehensiveness of our Fashionpedia ontology

Super-category	**class count**	*not sure*	*not on the list*
Length	15	12.79%	0.01%
Nickname	153	9.15 %	12.76%
Opening Type	10	32.69%	3.90%
Silhouettes	25	2.90%	0.27%
Tex finish, manu-tech	21	4.47%	1.34%
Textile Pattern	24	2.18%	5.30%
None-Textile Type	14	4.90%	4.07%
Neckline	25	9.57%	3.38%
Waistline	7	30.46%	0.17%

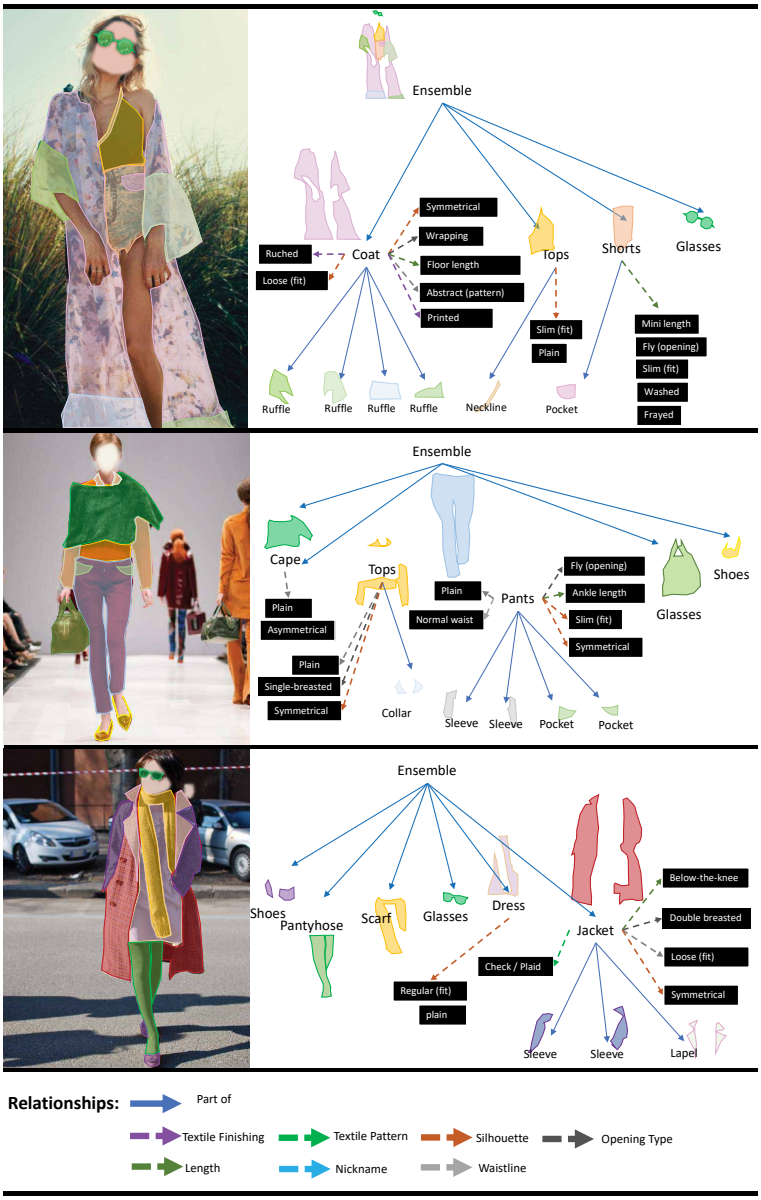


Fig. 11: Example images and annotations from our dataset: the images are annotated with both instance segmentation masks and fine-grained attributes (black boxes)

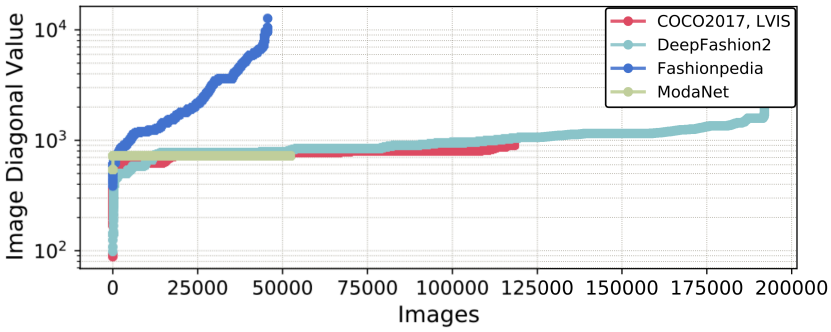


Fig. 12: Image size comparison among Fashionpedia, ModeNet, DeepFashion2, and COCO2017, LVIS. Only training images are shown. The Fashionpedia images has the most diverse resolutions. Note that COCO2017 and LVIS have higher resolution images for annotation. The distribution presented here are the publicly available photos

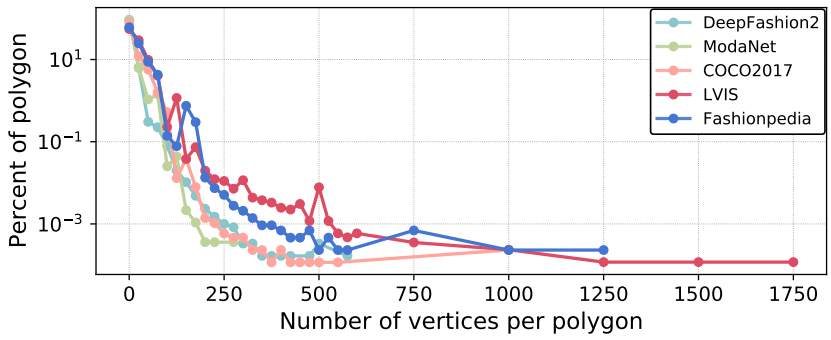


Fig. 13: The number of vertices per polygon. This represents the quality of masks and the efforts of annotators. Values in the x-axis were discretized for better visual effect. Y-axis is on log scale. Fashionpedia has the second widest range, next to LVIS

and make sure the clothing items were visible and annotatable in the images. The estimated image type breakdown is listed as follows: street style images (30% of the full dataset), celebrity events images (30%), runway show images (30%), and online shopping product images (10%). For gender distribution, the gender in 80% of images are female, and 20% of images are male.

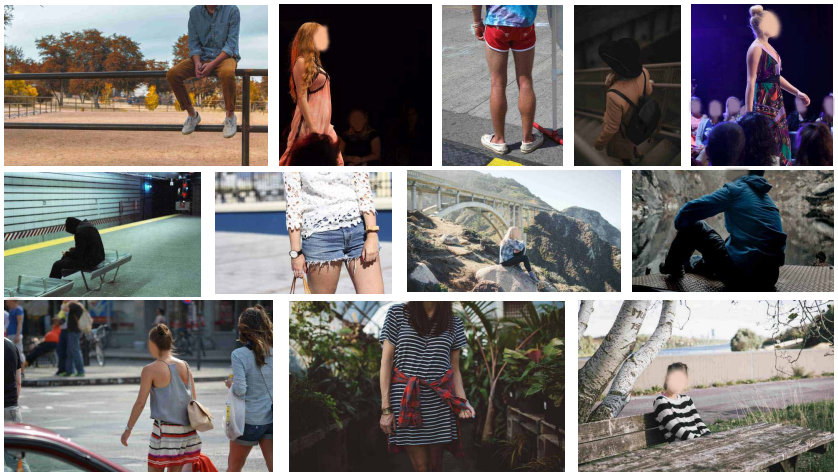


Fig. 14: Example of different type (people position/gesture, full/half shot, occlusion, scenes, garment types, etc.) of images in Fashionpedia dataset

We did aim to address the issue of photographic bias in the image collection process. Our dataset includes the images that are not centered, not full shot and with occlusion (see examples in Fig. 14). Furthermore, our focus is to identify clothing items, not to identify people during the image collection process.

Crowd workers and 10-day training for mask annotation. In the spirit of sharing the same apparel vocabulary for all the annotators, we prepared a detailed tutorial (with text descriptions and image examples) for each category and attributes in the Fashionpedia ontology (see Fig. 15 for an example). Before the official annotation process, we spent 10 days on training the 28 crowd workers for the following three main reasons.

First, some apparel categories are commonly referred as other names in general. For example, "top" is a general term for "shirt", "sweater", "t-shirt", "sweatshirt". Some annotators can mistakenly annotate a "shirt" as a "top". We need to train these workers so they have the same understanding of the proposed Fashionpedia ontology. Utilizing the prepared tutorials (see Fig. 15 for an example), we trained and educated annotators on how to distinguish among different apparel categories.

Second, there are fine-grained differences among apparel categories. For example, we observed that some workers initially had difficulty in understanding

Fig. 15: Annotation tutorial example for shirt and top

the difference among different garment parts, such as 'tassel' and 'fringe'. To help them understand the difference of these objects, we ask them to practice and identify more sample images before the annotation process. Fig. 16 shows our tutorials for these two categories. We specifically shows some correct and wrong examples of annotations.

Third, we ask for the quality of annotations. In particular, we ask the annotators to follow the contours of garments in the images as closely as possible. The polygon annotation process is monitored and verified for a few days before the workers started the actual annotation process.

Quality control of debatable apparel categories. During the annotation process, we allow annotators to ask questions about the uncertain categories. Two fashion experts monitored the annotation process by answering questions, checking the annotation quality, and providing weekly feedback to annotators.

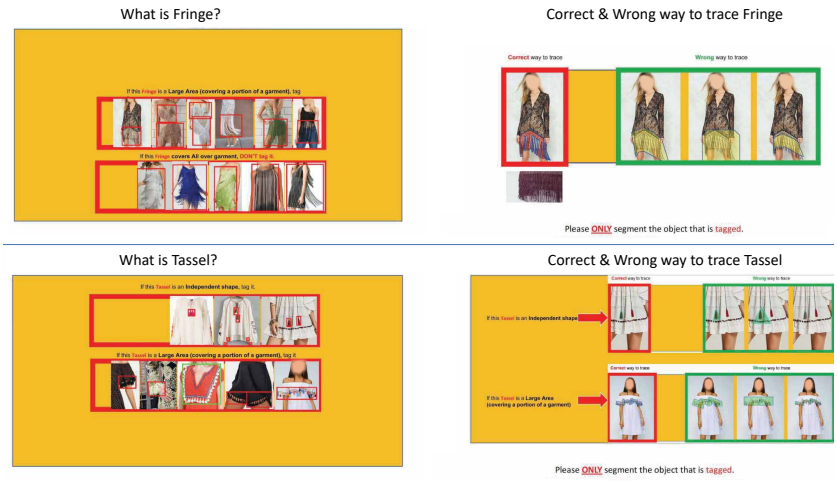


Fig. 16: Annotation tutorial for fringe and tassel

Instead of asking annotators to rate their confidence level of each segmentation mask, we asked them to send back all the uncertain masks to us during the annotation. The same two fashion experts made the final judgement and gave the feedback to the workers on these debatable or unsure fashion categories. Some examples of debatable or fuzzy fashion items that we have documented can be found in Figure 17.

1.5 Discussion

Does this dataset include the images or labels of previous datasets? We only include the previous datasets for comparison. Our dataset doesn't intentionally use any images or labels from previous datasets. All the images and labels from Fashionpedia are newly collected and annotated.

Who were the fashion experts annotating localized attributes in Fashionpedia dataset? The fashion experts are the 15 fashion graduate students that we recruited from one of the top fashion design institutes. For double-blind policy, we cannot mention the name of the university. But we will release the name of this university and the collaborators from this university in the final version of this paper.

Instance segmentation v.s. semantic segmentation. We didn't conduct semantic segmentation experiments on our dataset for the following two reasons: 1) Although semantic segmentation is a useful task, we believe instance segmentation is more meaningful for fashion images. For example, if we need to distinguish the different shoe style of a fashion image containing 3 pair of different shoes, instance segmentation (Figure 18(a)) can help us distinguish each shoe separately. However, semantic segmentation (Figure 18(b)) will mix all the

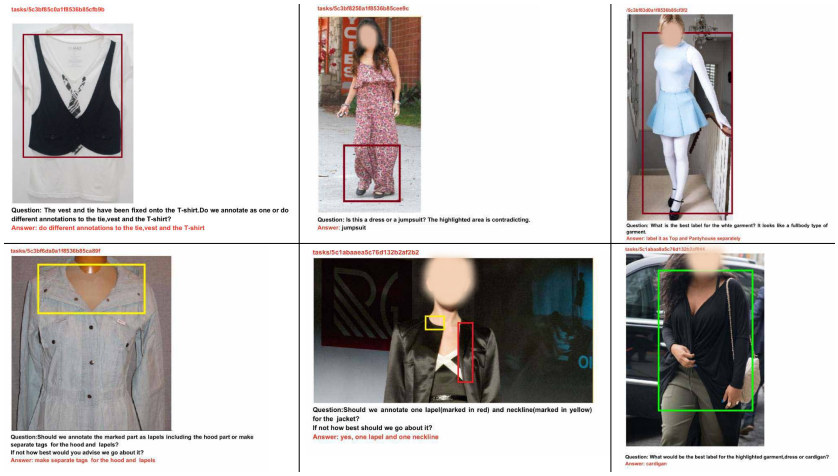


Fig. 17: Example of debatable fashion items in Fashionpedia dataset. The questions are asked by the crowdworkers. The answers are provided by two fashion experts

Fig. 18: Instance segmentation (left) and semantic segmentation (right)

shoe instances together. 2) Semantic segmentation is the sub-problem of instance segmentation. If we merge the same detected object class from our instance segmentation experimental result, it yields the results for semantic segmentation.

Which image has the most annotated masks? In Fashionpedia dataset, the maximum number of segmentation masks in an image is 74 (Fig. 19). We find that most of the masks are belonging to "rivets" (garment parts).

What's the difference between Fashionpedia and other fine-grained datasets like CUB-200? We propose to localize fine-grained attributes within segmentation masks of images. This is a novel task with real-world application to the best of our knowledge. The differences between Fashionpedia and CUB are as follows: 1) CUB uses keypoints as annotation to indicate different locations on birds, while Fashionpedia has segmentation masks of garments, garment parts, and accessories; 2) Fashionpedia attributes are associated with garment or garment part instances in images, whereas CUB provides global attributes, not associated with any specific keypoints.

iMat-Fashion Kaggle challenges To advance state-of-the-art of visual analysis of clothing, we hosted two kaggle challenges (imaterialist-fashion) on Kaggle in 2019 [1] and 2020 [2] respectively.

[1] https://www.kaggle.com/c/imaterialist-fashion-2019-FGVC6
[2] https://www.kaggle.com/c/imaterialist-fashion-2020-fgvc7

Fig. 19: The image with 74 masks in Fashionpedia dataset

References

1. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledge base (2014) 9
2. ZARA.com: Zara white leather flat ankle boots with top stitching size 5 bnwt, retrieved May 9, 2019 from https://www.zara.com 6