UNIVERSITAT DE BARCELONA

Object Recognition

Fashion parsing (segmentation)

Gerard Caravaca Ibáñez
gerard.caravaca.ibanez@estudiantat.upc.edu

24/4/2023

Contents

1 Introduction

Segmentation is the process of breaking up an image into different parts or segments according to some predetermined criteria, such as color, texture, or shape. Many computer vision tasks, including object detection and tracking, depend on it. In a range of applications, including medical imaging, surveillance, and autonomous driving, segmentation enables us to extract relevant information from a picture. There are numerous methods for segmentation, from straightforward thresholding to intricate algorithms based on machine learning and deep learning.

This report will deal with segmentation in a specific field such as fashion. A particular use of segmentation in the clothing and fashion industries is fashion semantic segmentation. It entails the method of segmenting a picture of a fashion item, such as clothing or accessories, depending on its semantic significance, such as recognizing the type of clothing or the material used in the product, Figure 1 shows an example of this task. To face this task we will analyse data from the fashionpedia dataset [2]. Subsequently, we will explore some methods based on deep learning to carry out the segmentation.

2 Dataset analysis

This section will explore the dataset used in this work. In order to be able to analyse the images used in depth, we will first give a brief description of the dataset and its purpose. Then, the distribution of the dataset will be analysed. Finally, the resolution of the images and the aspect ratio variation between images of the same dataset will be explored.

Fashionpedia is a distinctive database of clothing information created to unite the fields of fashion and computer vision. Its goal is to present a brand-new fine-grained segmentation challenge that blends the knowledge of the two disciplines. In order to make these methodologies effective in actual, real-world applications, the task's objective is to provide a uniform strategy to categorizing and segmenting certain garment qualities. For this purpose the dataset 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 48k everyday and celebrity event fashion images annotated with segmentation masks and their associated per-mask fine-grained attributes, built upon the Fashionpedia ontology.

For this task the second part of the dataset will be used.

Analyzing class distribution in segmentation datasets is crucial for ensuring that the resulting models are accurate and representative of the real-world data. In segmentation, class distribution refers to the distribution of pixels or regions across different classes or categories in the dataset. To explore this fact, it has been decided to count the number of times each class appears in the dataset obtaining Figure 2. In this figure it can be seen that the dataset is quite unbalanced. This is understandable considering the nature of the images. For example, the sleeves and shoes are very common garments used by everyone and will be repeated in most of the images. While, the umbrella will appear much less commonly. This difference in the distribution of classes makes the segmentation of less frequent classes in the dataset very complex.

On the other hand, distribution of height and weight can be quite important for segmentation model performance. If the dataset used to train the segmentation model is skewed towards particular height and weight ranges, the model may not function effectively on photos outside of those ranges. When segmenting shorter and wider objects, for instance, a segmentation model that has only been trained on images of tall and thin objects may have trouble doing so effectively. Because to the wide variation in item sizes and shapes in real-world photos, this can result in subpar performance. In case of fashionpedia dataset we can analyze the height and width distribution by explaining Table 1 and Figure 3. Table 1 shows some important statistics that make it clear that the difference in image size is notorious. In the case of Figure 3, it can be seen in more detail that most of the images are smaller than 4000x4000 pixels, but that the distribution is noticeable. However, this is about the original size of the images, for complexity, in this work we will use a maximum size of 1024x1024.

This analysis shows that, on the one hand, the distribution of classes is rather poor, which will make it difficult to obtain good results. In contrast, we have also shown that the distribution of image resolution is notorious which will help to generalise the model to all types of images, this fact is also demostrated in Figure 4.

3 Approach

To solve this task, after investigating some segmentation methods, I have decided to do fine-tunning on the OCRnet model proposed in [1]. The main contribution of this paper is the HRNet+OCR+SegFix system, which achieved the 1st place on the Cityscapes leaderboard by the ECCV2020 submission deadline, which is the main reason of choosing this method.

On the one hand, the OCR method employs a context scheme that is less complex than many other methods. It comprises three steps: soft object region extraction, computation of object region representation, and computation of object-contextual representation for each position. The method is implemented as 1x1 conv → BN → ReLU and utilizes a linear function (a 1x1 convolution) to predict the coarse segmentation, which is supervised with a pixel-wise cross-entropy loss. The OCR is obtained by aggregating all the object region representations with weights calculated based on the relationships between pixels and object regions (see Figure 5).

On the other hand, several backbone architectures have been tested in the paper. I decided to take HRNet-W18, which is a convolutional neural network architecture with a wide 18-layer network. It has a stride of 4, meaning that the convolutional filters are applied with a stride of 4 pixels, allowing for a more detailed feature extraction than a stride of 8. It is used as the basis for the OCR model and pre-trained on the Voc12Aug dataset. Even that more capable versions are available, version 18 layers has been chosen because of Colab's memory usage limit.

To implement this method, I have used the mmsegmentation repository. It is an open-source image segmentation toolbox based on PyTorch. It provides a variety of state-of-the-art semantic segmentation and instance segmentation models, including FCN, U-Net, PSPNet, DeepLabV3, Mask RCNN, and many more. You can find the repository on GitHub: mmsegmentation.

4 Experimentation

Having described the dataset and the model used, it only remains to experiment with different parameters and analyse the results. In this section, I will outline the process followed for fine-tuning the model described above. Consequently, I will show the results by testing the model in the validation/test split of the fashionpedia dataset. Finally, some other new approaches will be commented.

When training the model we use the following list of parameters. I have trained the model with 0.01 lerning rate with decay according to the epoch. I have also used a batch size of 8 and an image size of 256x256. On the other hand, random flip has been applied as data augmentation. Finally the model has been trained for 20000 epochs. All other parameters of the model have been left default. It should be noted that these parameters may not be the most efficient for the model and the dataset. However, they were the ones that gave the best results given the resources available.

Once the model has been trained with the parameters specified above, it is time to evaluate the model on validation and test data. The segmentation results shown in Table 2 are evaluated based on two metrics, IoU (Intersection over Union) and accuracy (Acc), for different clothing classes. Looking at the results, we can see that the background class has a very high IoU of 97.06% and accuracy of 99.14%, which is expected since the background is typically the largest class in an image. On the other hand, some of the clothing classes such as cardigan, vest, jumpsuit, cape, and various accessories have IoU and accuracy scores of 0.0, indicating that the model failed to accurately segment these classes. Among the clothing classes with non-zero IoU scores, we can see that the segmentation performance is generally poor with mIoU (mean IoU) of 14.93% and mAcc (mean accuracy) of 19.01%. However, some classes such as pants, dress, and sleeve have relatively higher IoU and accuracy scores, indicating that the model was able to segment these classes more accurately. As a summary of the results, it is clear that the classes that are repeated most often in the dataset are the ones that are best predicted by the model. This can be seen in some examples set out in: Figure 6, Figure 7 and Figure 8.

It is worth noting that in the last week a model has emerged that far surpasses any fine-tunnnig I have been able to test. It is Segment Anything Model (SAM) [3], a promptable segmentation system with zero-shot generalization to unfamiliar objects and images, without the need for additional training. This Meta-trained model is capable of segmenting almost any type of object. You can see some promising results in fashionpedia inference in Figure 9.

5 Conclusions

To summarise, The main contribution of this work is the contribution of a fine-tuned model to face the task of garment segmentation in the field of fashion. The model tested is ocrnet using hrnet as backbone architecture pre-trained on Voc12Aug dataset. After fine-tuning the model with various parameter settings, results of around 19% mAcc have been achieved. These results leave room for improvement. Some points I leave for the improvement of the results is the possible improvement of the dataset. This is because it has been shown that the dataset is completely unbalanced and this causes the model to learn worse. Another possible improvement may be in the size of the pre-trained model selected, as it has not been possible to test larger models because of the memory limit. Finally, a new approach has been proposed based on the application of a model launched last week by the company Meta. As I have shown, this segmentation model promises to be a revolution in the field.

6 Appendix

6.1 Introduction

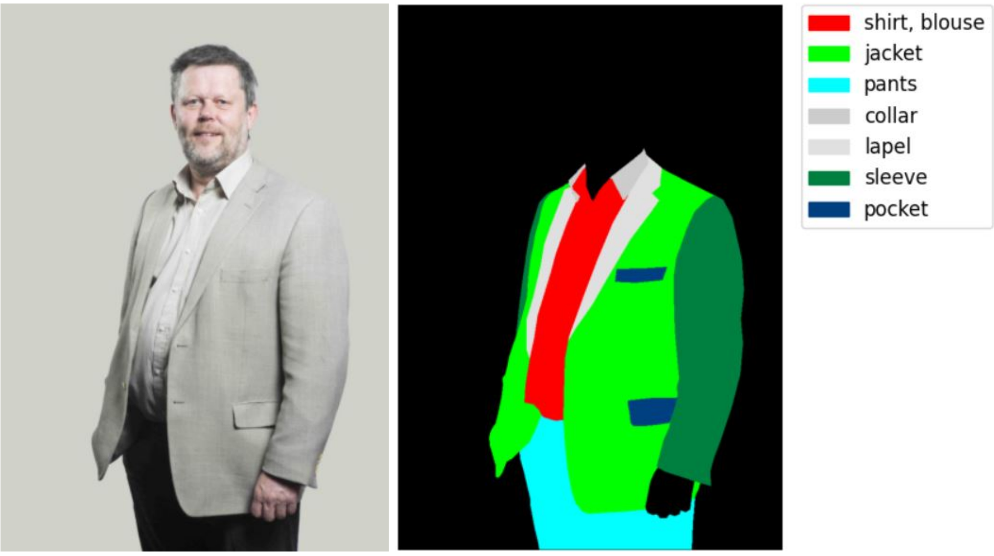

Fig. 1: Examples of segmentation task in fashion.

6.2 Dataset analysis

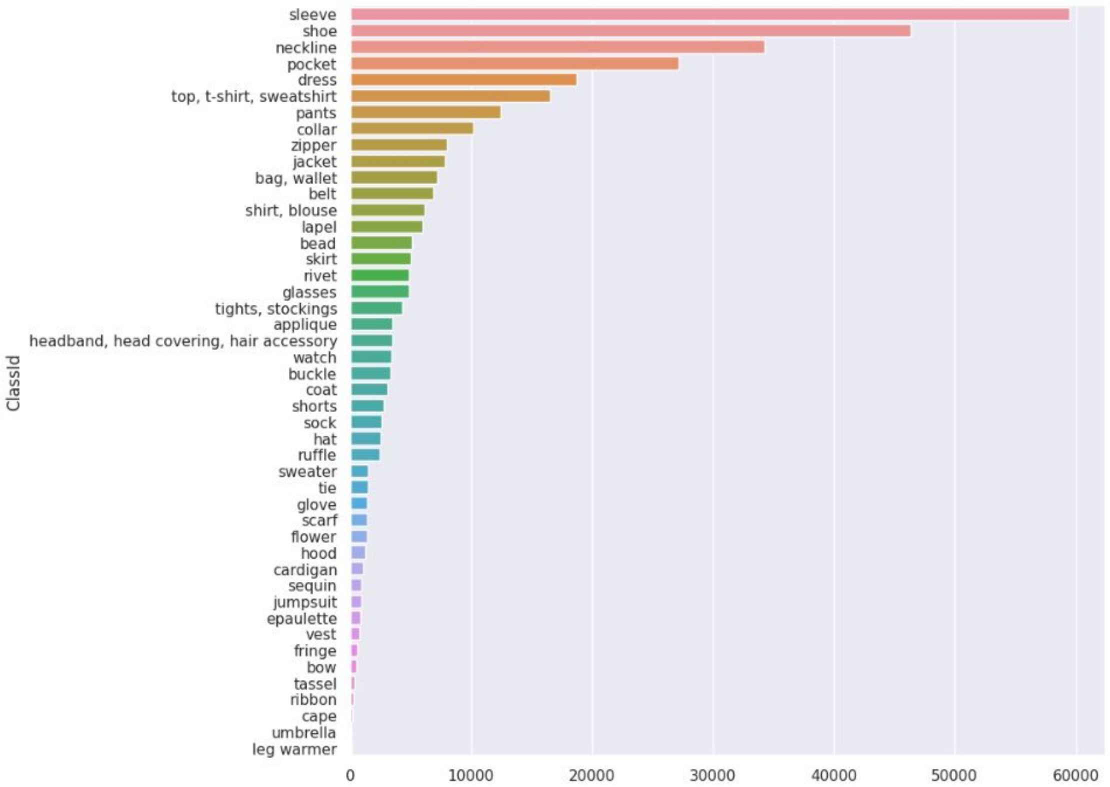

Fig. 2: Class distribution of fashionpedia dataset.

	Height	Width
Max	8688.000000	10717.00000
Min	296.000000	151.00000
Mean	2235.524021	1761.97925

Tab. 1: Height and width statistics of original fashionpedia images.

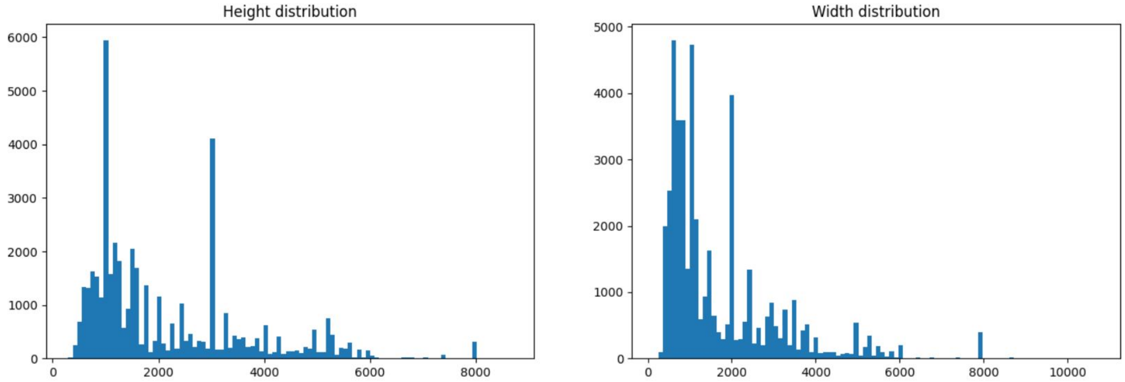

Fig. 3: Height and width distribution of original fashionpedia images.

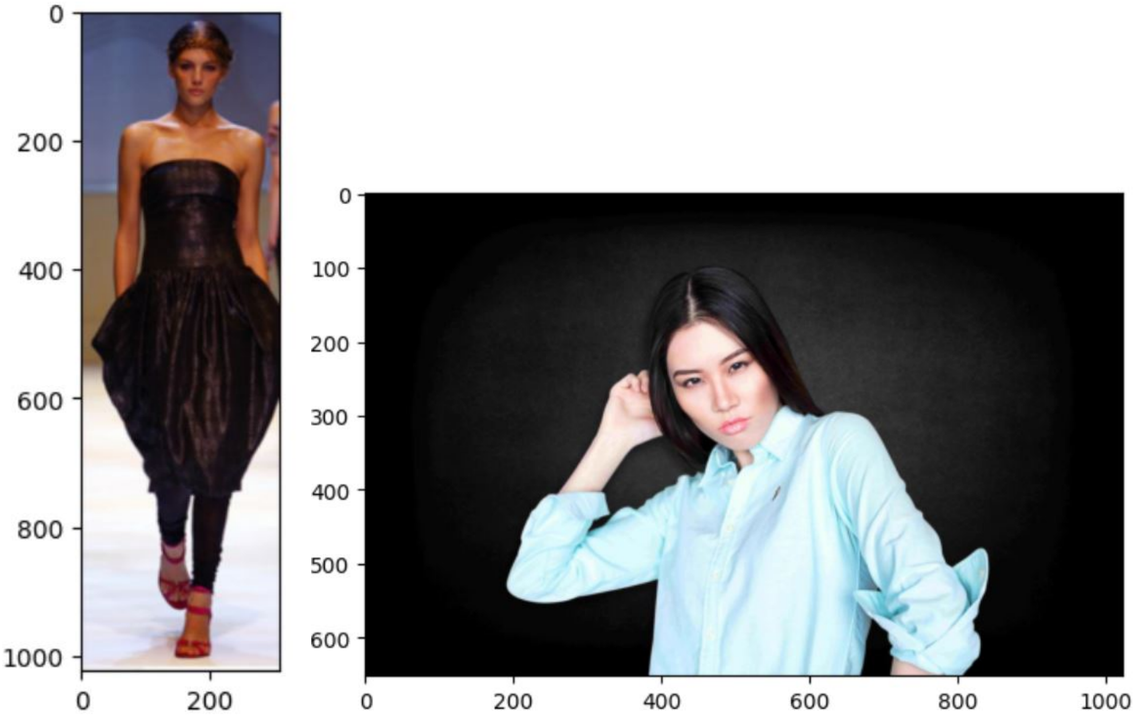

Fig. 4: Examples of images with different resolutions.

6.3 OCRnet

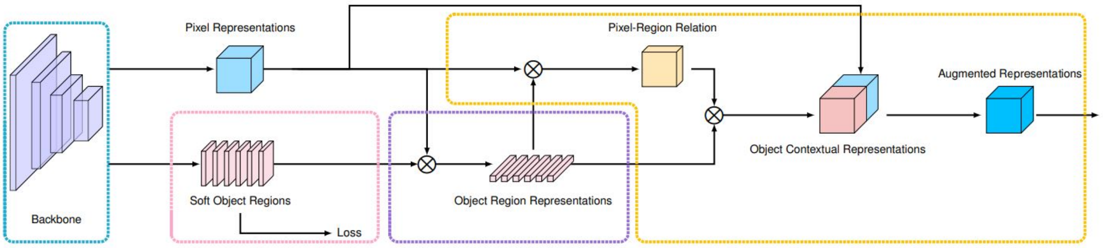
Pixel Representations
Pixel-Region Relation
Augmented Representations
Object Contextual Representations
Soft Object Regions
Object Region Representations
Backbone
Loss

Fig. 5: OCR's pipeline.

6.4 Experimentation

Class	**IoU**	**Acc**
background	97.06	99.14
shirt, blouse	26.63	36.85
top, t-shirt, sweatshirt	50.08	69.91
sweater	18.07	23.07
cardigan	0.0	0.0
jacket	33.03	47.57
vest	0.0	0.0
pants	58.17	75.78
shorts	43.62	58.49
skirt	37.83	48.31
coat	31.67	46.56
dress	64.65	83.19
jumpsuit	5.39	8.91
cape	0.0	0.0
glasses	1.1	1.1
hat	19.53	22.98
headband, head covering, hair accessory	0.0	0.0
tie	0.0	0.0
glove	0.0	0.0
watch	0.0	0.0
belt	0.53	0.54
leg warmer	0.0	0.0
tights, stockings	23.78	28.77
sock	0.0	0.0
shoe	47.61	60.28
bag, wallet	27.98	38.74
scarf	0.57	0.59
umbrella	0.0	0.0
hood	0.84	0.85
collar	11.34	14.35
lapel	19.12	25.64
epaulette	0.0	0.0
sleeve	57.06	73.48
pocket	1.85	1.93
neckline	9.77	11.37
buckle	0.0	0.0
zipper	0.0	0.0
applique	0.0	0.0
bead	0.0	0.0
bow	0.0	0.0
flower	0.0	0.0
fringe	0.0	0.0
ribbon	0.0	0.0
rivet	0.0	0.0
ruffle	14.27	14.88
sequin	0.0	0.0
tassel	0.0	0.0
aAcc	**mIoU**	**mAcc**
93.7	14.93	19.01

Tab. 2: IoU and Acc scores for different clothing classes

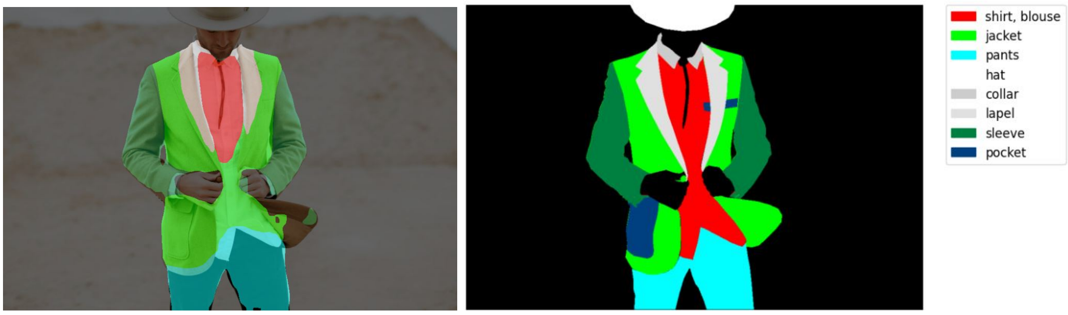

Fig. 6: Example 1 of inference of the model (segmentation in left, ground truth in right).

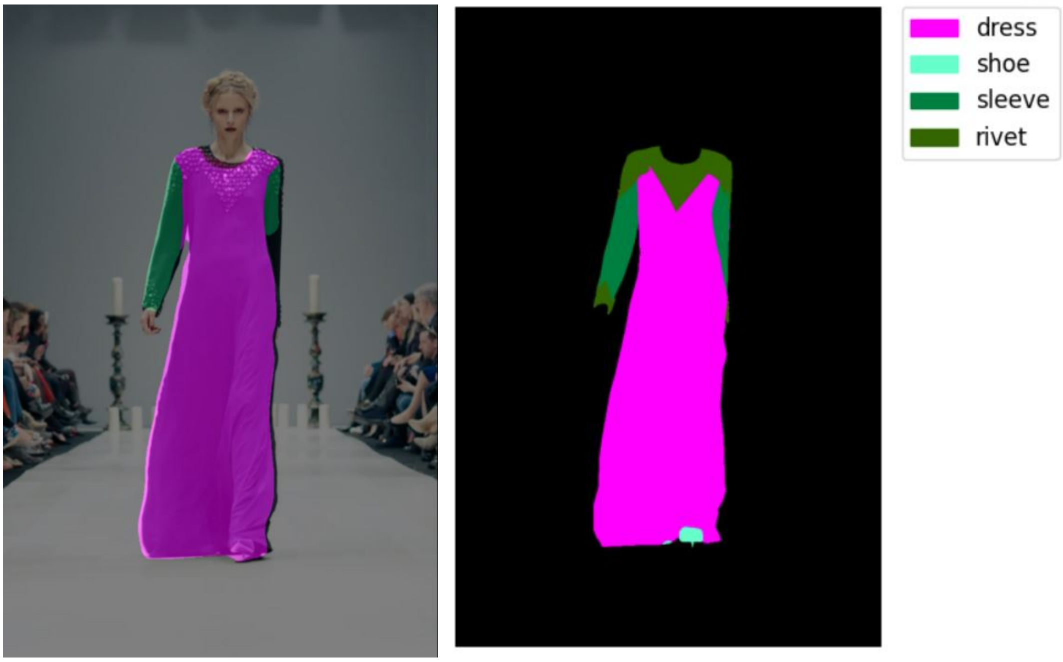

Fig. 7: Example 2 of inference of the model (segmentation in left, ground truth in right).

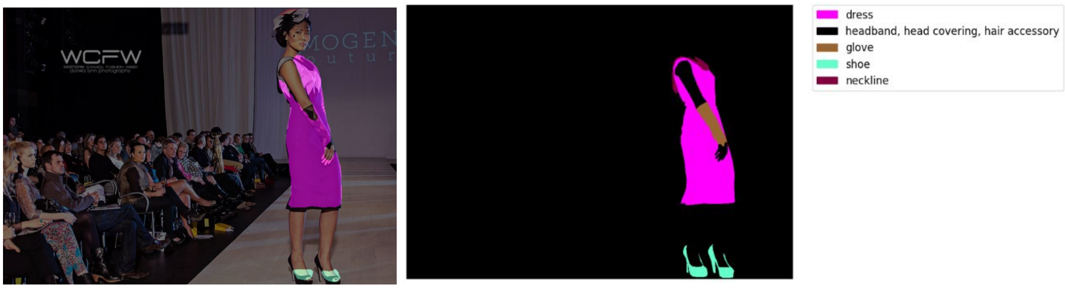

Fig. 8: Example 3 of inference of the model (segmentation in left, ground truth in right).

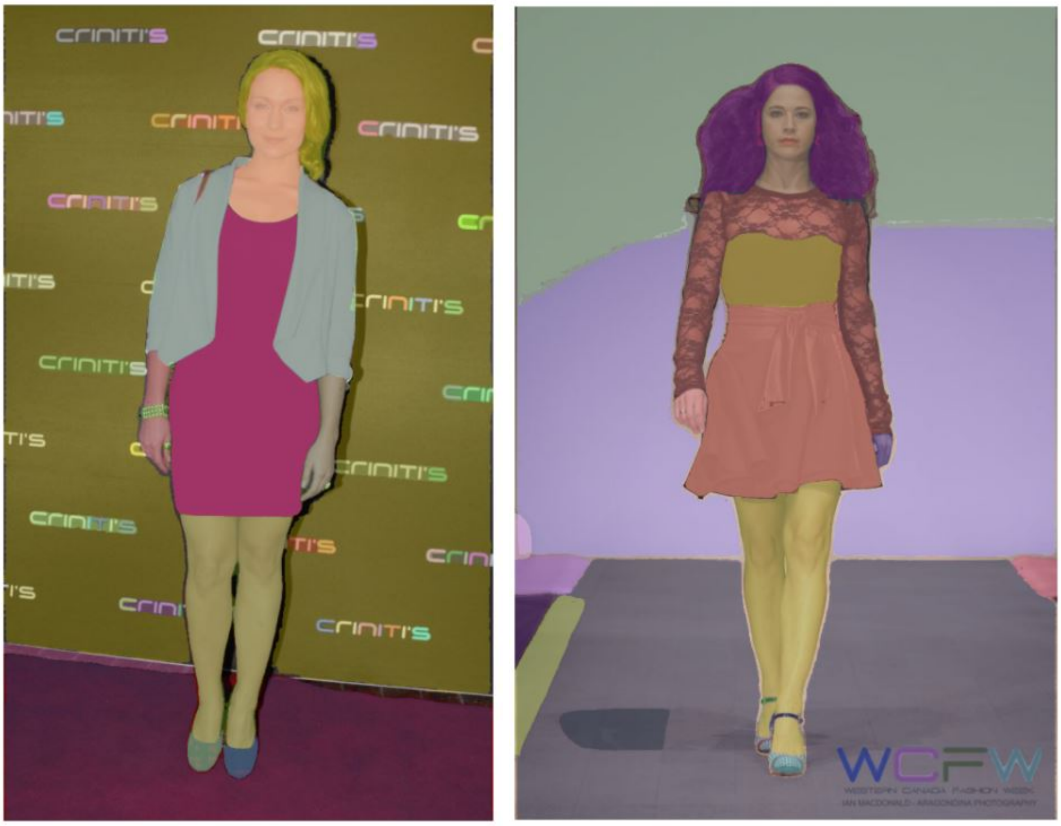

Fig. 9: Inference of SAM model in fashionpedia.

References

[1] Yuhui Yuan et al. “Segmentation transformer: Object-contextual representations for semantic segmentation”. In: *arXiv preprint arXiv:1909.11065* (2019).

[2] Menglin Jia et al. “Fashionpedia: Ontology, Segmentation, and an Attribute Localization Dataset”. In: *CoRR* abs/2004.12276 (2020). arXiv: 2004.12276. url: https://arxiv.org/abs/2004.12276.

[3] Alexander Kirillov et al. “Segment anything”. In: *arXiv preprint arXiv:2304.02643* (2023).