Clothing Detection and Segmentation

Apollinaria Chernikova
Innopolis University
Innopolis, Russia
a.chernikova@innopolis.university

Egor Machnev
Innopolis University
Innopolis, Russia
e.machnev@innopolis.university

I. Introduction

The rapid evolution of computer vision techniques has enabled transformative applications in diverse industries, including fashion. Clothing detection and segmentation play a pivotal role in automating tasks such as cataloging, personalized recommendations, and virtual try-on systems [1]. These capabilities are foundational for modern retail platforms aiming to deliver enhanced user experiences and operational efficiency.

There are different types of segmentations. In our work we want to explore basic approaches for instance segmentation of clothes.

II. Related Work

The field of clothing detection and segmentation has historically received less attention compared to other object detection and segmentation tasks. Researchers often prioritize domains like autonomous driving or medical imaging due to their broader industrial applications. Despite this, several notable works have made significant contributions to advancing clothing-specific detection and segmentation, which serve as the foundation for this study.

A. DeepFashion2 and Related Datasets

Datasets are the backbone of deep learning research in computer vision, and the clothing detection domain is no exception. The DeepFashion2 dataset has played a pivotal role in this field by providing large-scale, diverse annotations for clothing items, including bounding boxes, segmentation masks, pose information, and attribute labels [2]. The dataset's comprehensive coverage of various clothing types, styles, poses, and occlusions makes it an invaluable benchmark for evaluating detection and segmentation models.

Complementary datasets such as ModaNet [3] and Fashionpedia [4] have addressed specific challenges in this domain. ModaNet emphasizes fine-grained clothing part segmentation and provides high-quality annotations tailored for fashion analytics. On the other hand, Fashionpedia combines segmentation with ontology-based attribute annotations, offering hierarchical labels that integrate well with high-level understanding tasks. These datasets collectively enrich the landscape of clothing detection and segmentation research, allowing for robust evaluation across diverse scenarios.

B. Advanced in Model Architectures

In recent years, researchers have made significant strides in designing architectures for clothing detection and segmentation. Many studies have built upon existing instance segmentation frameworks, with Feature Pyramid Networks (FPN) serving as a cornerstone in state-of-the-art models [5], [6]. FPNs enable efficient multi-scale feature extraction, which is particularly beneficial in scenarios with variable clothing sizes and complex patterns.

Among the most prominent architectures, **Mask R-CNN** equipped with FPN has emerged as a leading approach in the field. This model combines region-based detection with pixel-wise segmentation, achieving impressive results on benchmarks such as DeepFashion2 and ModaNet [6]. Its ability to adapt to challenging scenarios like occlusions and overlapping garments has made it the standard for clothing segmentation tasks.

While many researchers explore custom architectures for enhanced performance [7], the integration of robust components like FPN and ResNet backbones remains a consistent theme. These elements not only improve segmentation accuracy but also ensure scalability across datasets and real-world applications.

III. METHODOLOGY

A. *Dataset*

For this study, we utilized the **DeepFashion2** dataset [2], one of the most comprehensive datasets for clothing detection and segmentation. The dataset contains 191K images for training and 32K images for validation. Each image is labeled with bounding boxes, segmentation masks, and clothing attributes. The dataset includes clothing across 13 main categories and covers a wide range of scenarios, including occluded items, varying poses, and diverse lighting conditions.

.

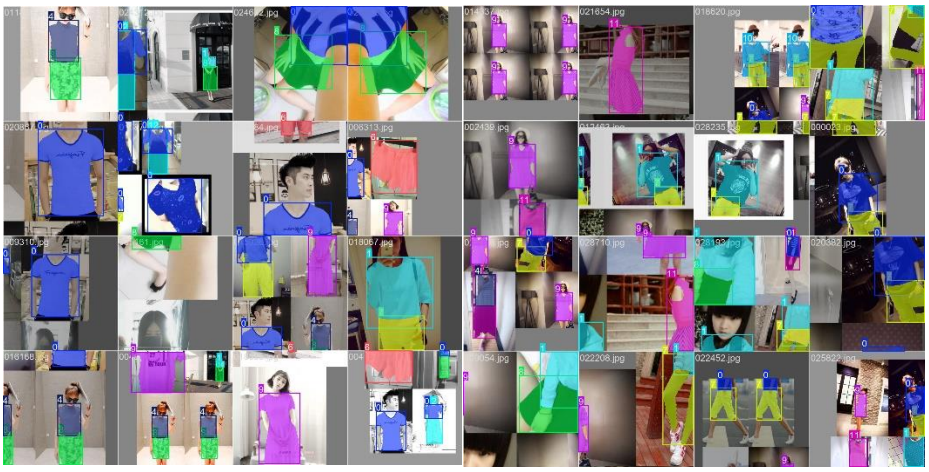

Fig I. DEEPFASHION2 IMAGES EXAMPLE

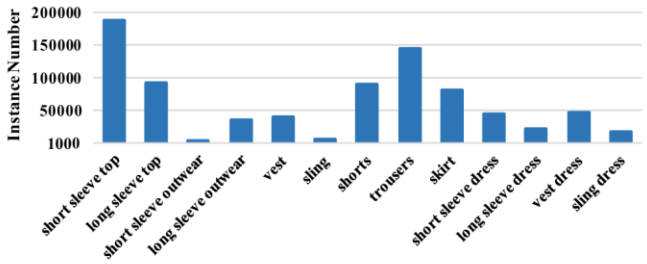

Fig II. DEEPFASHION2 CATEGORIES DISTRIBUTION

B. *Model Architecture*

Given the dual objectives of detection and segmentation, we chose a combination of the **Mask R-CNN** [8] and **YOLOv8** [9] architectures. The choice of these architectures was based on their strengths in addressing the specific requirements of this study.

Mask R-CNN [8] was chosen for its well-established capabilities in instance segmentation, particularly its ability to generate accurate pixel-wise segmentation masks. With its region proposal network (RPN) and Feature Pyramid Network (FPN), Mask R-CNN effectively handles multi-scale detection and segmentation tasks. This model was ideal for achieving high accuracy in segmentation, especially in scenarios where detailed masks were required to differentiate overlapping or occluded clothing items.

YOLOv8 [9], on the other hand, was selected for its efficiency in detection and its ability to operate in real-time scenarios. Unlike Mask R-CNN, YOLOv8 emphasizes speed and computational efficiency while maintaining reasonable accuracy. Although primarily a detection model, the segmentation extension of YOLOv8 allowed for basic segmentation capabilities, which were leveraged in this study to compare its performance with Mask R-CNN. YOLOv8's lightweight nature made it a suitable candidate for applications where processing speed was critical.

Both models were fine-tuned using pre-trained weights to expedite the training process and improve performance. Specifically, the **Detectron2** framework was used for training the Mask R-CNN model. Detectron2 is a highly efficient implementation of several state-of-the-art object detection algorithms, including Mask R-CNN, and is built on top of PyTorch. The pre-trained weights for Detectron2 were obtained from a model trained on large-scale datasets like COCO, which significantly accelerated convergence during training and improved model performance on the DeepFashion2 dataset.

Similarly, YOLOv8 was initialized with pre-trained weights, leveraging a model trained on large object detection datasets like COCO to enhance learning on fashion-specific tasks. Fine-tuning these models on the DeepFashion2 dataset enabled them to adapt to the nuances of clothing detection and segmentation.

C. Hardware Resources

All training experiments were conducted using the maximum computational resources available to us. The primary hardware used was the **NVIDIA Tesla P100 GPU**, which offers **16 GB of GPU memory** and **30 GB of system memory (RAM)**.

IV. GitHub link

Link to our GitHub repository with all experiments and evaluations: https://github.com/ApollyCh/cloth-detection

V. Experiments and Evaluation

A. Training Pipeline

For the training process, we selected three different model configurations: **Mask R-CNN ResNet-50 3x**, **Mask R-CNN ResNet-101 3x**, and **YOLOv8n** for segmentation tasks. Given the complex nature of clothing detection and segmentation, we performed multiple runs with varying hyperparameters for the Mask R-CNN models to identify the optimal configuration.

The parameters for these models were chosen particularly from the **DeepFashion2** dataset findings [2], and the computational resources available to us.

Hyperparameters	Value
Base Learning Rate (LR)	0.002
Number of Iterations	5000
Number of Warm-up Iterations	500
Gamma	0.1
Images per Batch	8
Batch Size per Image	256

Fig III. Mask-RCNN Training Parameters

For the Mask R-CNN models, we experimented with ResNet-50 and ResNet-101 backbones. The ResNet-50 model was used for its relatively lower computational cost and quicker training time, while ResNet-101 was chosen for its deeper architecture, offering potential improvements in segmentation accuracy, especially for fine-grained details in clothing items.

In the context of our experiments, we initially planned to train the **YOLO** model in addition to Mask R-CNN models. However, due to the limited computational resources available and the large size of the DeepFashion2 dataset, training YOLO proved to be impractical. The training process for YOLO was excessively slow and took an unreasonably long time to complete. This raised concerns about the feasibility of using YOLO for such large datasets, particularly in a real-world scenario where fast training and inference are required.

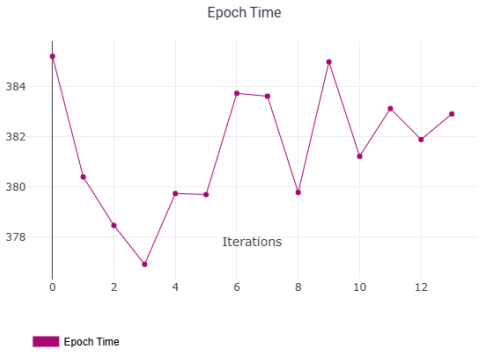

Fig IV. TIME PER EPOCH FOR YOLO MODEL

Hence, given the constraints, we made the decision to shift our focus entirely to the Mask R-CNN architectures. This decision allowed us to dedicate computational resources more effectively to the Mask R-CNN models, which demonstrated good performance and faster convergence compared to YOLO under the given circumstances.

B. Evaluation Pipeline

For the evaluation of the models, we used metrics that are widely accepted in the field of instance segmentation to assess the performance of the trained models. Specifically, we focused on **Average Precision (AP)** and **Average Recall (AR)** as our primary evaluation metrics. These metrics provide a clear and interpretable measure of a model's effectiveness in detecting and segmenting individual objects [2], [5], [6].

VI. ANALYSIS AND OBSERVATIONS

During the experiments, both models—Mask R-CNN ResNet-50 and Mask R-CNN ResNet-101—demonstrated strong performance. The total loss decreased significantly in the initial epochs and continued to improve steadily throughout training. This indicates that both models were effectively learning from the dataset and converging towards optimal performance.

Fig V. LOSSES COMPARISON

The primary difference observed between the two architectures was the training speed. Due to its smaller number of layers, the ResNet-50 backbone trained faster than the ResNet-101 backbone. This made ResNet-50 more resource-efficient, which could be advantageous in scenarios where computational resources or training time are limited.

As previously mentioned, AP and AR were the primary metrics used to evaluate the models. Additionally, we evaluated AP for individual classes to examine the performance of the models across the various categories in the DeepFashion2 dataset. This analysis was crucial because the dataset is inherently unbalanced, with some classes being underrepresented due to the uneven stratification of clothing categories. This class imbalance can introduce bias into the model's predictions, with the risk of overfitting to dominant classes while underperformed on less frequent ones.

Models	**AP**	**AP50**	**AP75**	**AR**
mask-rcnn-r-50-fpn-3x	0.412	0.559	0.481	0.648
mask-rcnn-r-101-fpn-3x	0.434	0.582	0.507	0.656

	AP			**AP**			**AP**	
Category	r-50	r-101	**Category**	r-50	r-101	**Category**	r-50	r-101
Short sleeved shirt	0.70	0.708	Long sleeved shirt	0.564	0.596	Short sleeved outwear	0.00	0.02
Long sleeved outwear	0.34	0.377	Vest	0.433	0.434	Sling	0.03	0.03
Shorts	0.528	0.532	Trousers	0.646	0.668	Skirt	0.562	0.578
Short sleeved dress	0.478	0.499	Long sleeved dress	0.310	0.355	Vest dress	0.499	0.532
Sling dress	0.266	0.297						

Fig VI. EVALUATION METRICS COMPARISON

Analyzing the evaluation metrics, it is evident that the Mask R-CNN ResNet-101 model outperforms the ResNet-50 model. This outcome aligns with our initial expectations, as the deeper architecture of ResNet-101 allows it to learn more complex patterns and features from the data. However, the improvement is marginal, with ResNet-101 achieving slightly higher values for AP and AR metrics compared to ResNet-50.

When examining the results across categories, noticeable differences between the two models become apparent. As shown in the images in Appendix, ResNet-101 produces more precise segmentation boundaries. The model is better at accurately delineating the edges of clothing items, even in cases where occlusions or overlaps are present. In contrast, ResNet-50 occasionally struggles with defining clear boundaries, particularly in complex scenes with multiple overlapping garments.

VII. CONCLUSION

In conclusion, we established our own baseline for further research in the domain of clothing detection and segmentation. Through our experiments, we confirmed that the Mask R-CNN architecture is highly suitable for tackling tasks involving large-scale datasets like DeepFashion2, even in scenarios with limited computational resources.

Both models, ResNet-50 and ResNet-101, demonstrated solid performance, successfully detecting and segmenting clothing items across a diverse range of categories and conditions. ResNet-101 provides slightly better segmentation accuracy and produces more precise boundaries, making it a better choice for applications where segmentation quality is critical. However, ResNet-50, with its faster training time and lower computational requirements, is more suitable for scenarios where speed and efficiency are prioritized alongside good performance.

APPENDIX

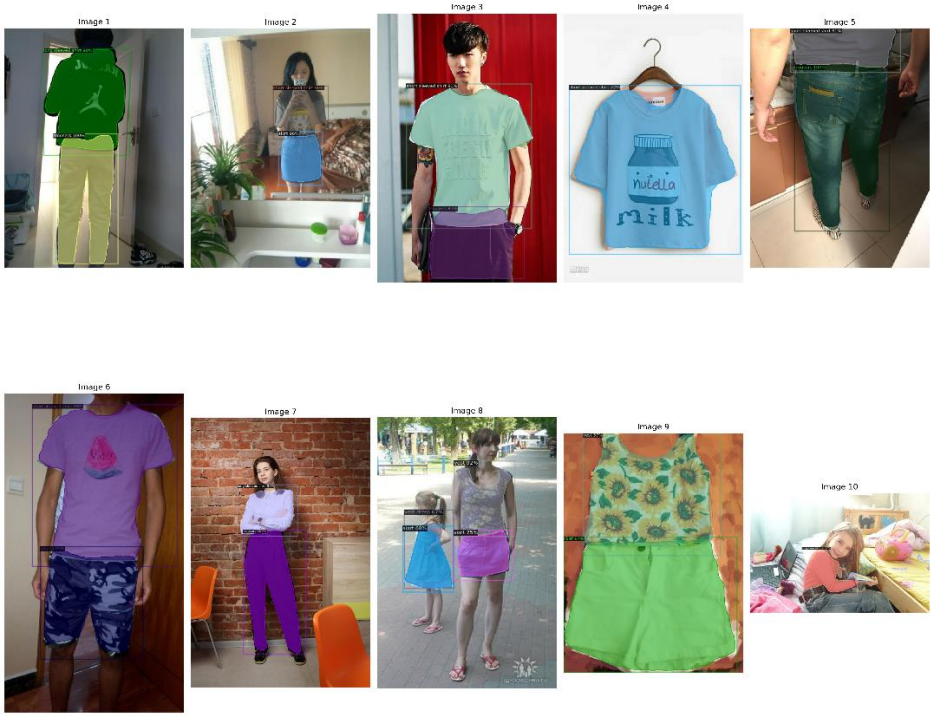

Fig VII. MASK-RCNN-R50 IMAGES EVALUATION

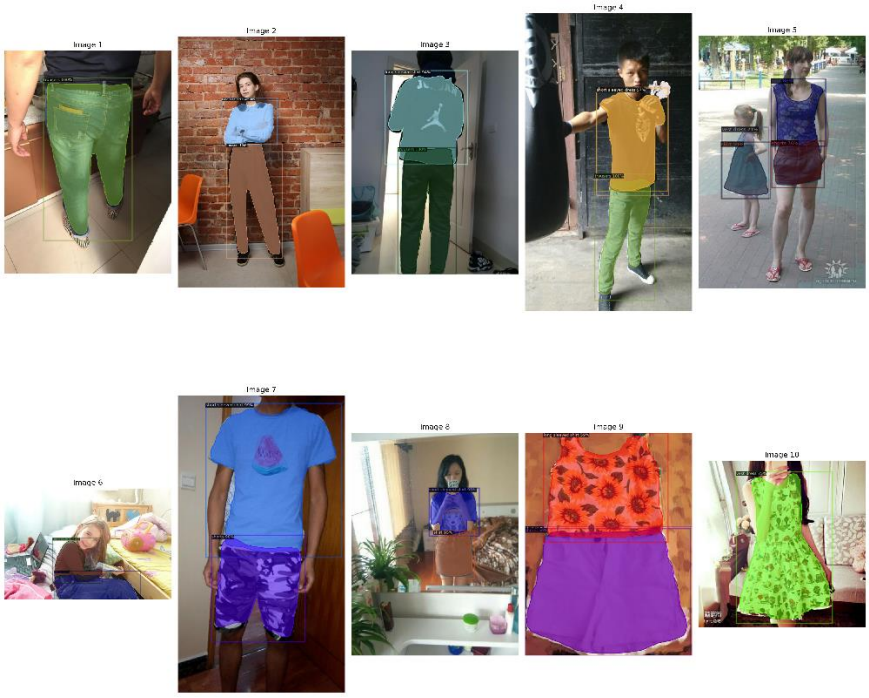

Fig VIII. MASK-RCNN-R101 IMAGES EVALUATION

VIII. REFERENCE

[1] Guangyu Tang, Feng Yu, Huiyin Li, Yankang Shi, Li Liu, Tao Peng, Xinrong Hu, Minghua Jiang, ClothSeg: semantic segmentation network with feature projection for clothing parsing, Journal of Visual Communication and Image Representation, Volume 97, 2023, 103980, ISSN 1047-3203, https://doi.org/10.1016/j.jvcir.2023.103980.

[2] Yuying Ge and Ruimao Zhang and Lingyun Wu and Xiaogang Wang and Xiaoou Tang and Ping Luo , DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images, 2019, https://arxiv.org/abs/1901.07973

[3] Shuai Zheng and Fan Yang and M. Hadi Kiapour and Robinson Piramuthu, ModaNet: A Large-Scale Street Fashion Dataset with Polygon Annotations, 2019, https://arxiv.org/abs/1807.01394

[4] Menglin Jia and Mengyun Shi and Mikhail Sirotenko and Yin Cui and Claire Cardie and Bharath Hariharan and Hartwig Adam and Serge Belongie, Fashionpedia: Ontology, Segmentation, and an Attribute Localization Dataset, 2020, https://arxiv.org/abs/2004.12276

[5] Tsung-Yi Lin and Piotr Dollár and Ross Girshick and Kaiming He and Bharath Hariharan and Serge Belongie, Feature Pyramid Networks for Object Detection, 2017, https://arxiv.org/abs/1612.03144

[6] Martinsson, John & Mogren, Olof. (2019). Semantic Segmentation of Fashion Images Using Feature Pyramid Networks. 3133-3136. 10.1109/ICCVW.2019.00382.

[7] Shilin Xu and Xiangtai Li and Jingbo Wang and Guangliang Cheng and Yunhai Tong and Dacheng Tao, Fashionformer: A simple, Effective and Unified Baseline for Human Fashion Segmentation and Recognition, 2022, https://arxiv.org/abs/2204.04654

[8] Kaiming He and Georgia Gkioxari and Piotr Dollár and Ross Girshick, Mask R-CNN, 2018, https://arxiv.org/abs/1703.06870

[9] R. Varghese and S. M., "YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness," *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, Chennai, India, 2024, pp. 1-6, doi: 10.1109/ADICS58448.2024.10533619.

[10] Tsung-Yi Lin and Michael Maire and Serge Belongie and Lubomir Bourdev and Ross Girshick and James Hays and Pietro Perona and Deva Ramanan and C. Lawrence Zitnick and Piotr Dollár, Microsoft COCO: Common Objects in Context, 2015, https://arxiv.org/abs/1405.0312