Where are my clothes? A multi-level approach for evaluating deep instance segmentation architectures on fashion images

Warren Jouanneau[1,2], Aurélie Bugeau[1], Marc Palyart[2], Nicolas Papadakis[3], Laurent Vézard[2]
[1]Univ. Bordeaux, Bordeaux INP, CNRS, LaBRI, UMR 5800,F-33400 Talence, France
[2]Lectra, F-33610 Cestas, France
[3]Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251,F-33400 Talence, France
warren.jouanneau@u-bordeaux.fr

Abstract

In this paper we present an extensive evaluation of instance segmentation in the context of images containing clothes. We propose a multi level evaluation that completes the classical overlapping criteria given by IoU. In particular, we quantify both the contour and color content accuracy of the the predicted segmentation masks. We demonstrate that the proposed evaluation framework is relevant to obtain meaningful insights on models performance through experiments conducted on five state of the art instance segmentation methods.

1. Introduction

Clothes segmentation is the cornerstone of many image processing tasks in the fashion industry. Although segmentation is already useful by itself to isolate a garment from an outfit for display purposes, it is predominantly used as a pre-processing step for numerous applications: virtual try on for obtaining a source apparel [14], visual semantic embedding to obtain the products from an outfit [1], color applications such as color harmony [9].

In addition to these applications where segmentation is already well grounded, other use cases that solely rely on detection as a pre-processing step could be improved with segmentation. For example, clothes retrieval is traditionally performed with detection [12] but segmentation has proven a viable alternative [2].

Over the years, a wide range of deep segmentation models have been proposed [10, 11, 4, 5, 3]. Selecting the best one for a specific use case can end up being a daunting task. The current de facto standard approach for evaluating a segmentation architecture is the mean average precision (mAP) which is based on the intersection over the union (IoU) also known as the Jaccard index. Its main strength lies in its ability to sum up the performance with a unique metric. However this approach suffers from two major limitations. It does not capture well the quality of the contour and does not take into account the content associated with the identified masks.

These drawbacks raise the need for a broader evaluation method including different aspects that are crucial in the fashion context. An evaluation protocol giving insights on the quality of predicted masks would be of great interest to benchmark methods, finding applications in machine learning model training, model selection and model drift analysis in production. Therefore, for evaluating segmentation architectures, we propose in this paper a multi-level approach that relies on three levels: global, contour and content.

The remainder of the paper is organized as follow. In Section 2 we present in details the limitations of the mainstream approach for evaluating the performance of segmentation architectures. We offer an alternative by introducing in Section 3 our multi-level evaluation approach. Finally in Section 4 we use the proposed approach to compare and evaluate the performance of existing state-of-the-art deep segmentation architectures on fashion images.

2. mAP for instance segmentation evaluation

The dominant approach for evaluating instance segmentation methods is the mAP. The mAP is the mean of the average precision AP_c computed per class (c) over all the possible classes (C):

$$mAP_\alpha = \sum_{c \in C} \frac{AP_{c,\alpha}}{|C|}, \tag{1}$$

where α is a threshold used to discriminate true and false predictions needed to compute the precision-recall curve. We recall that the average precision (AP) is the area under the precision-recall curve.

In order to compute AP_α an underlying metric is used. For object detection or instance segmentation, the metric

Figure 1: Hand generated T-shirt masks (G_i) that all have a IoU score of approximately 0.61 with ground truth (GT). Each G_i contains the computed IoU, BJ, $sEMD$ with the ground truth.

used to fill this role is the intersection over the union (IoU) between a predicted mask $\hat{Y}$ and a ground truth mask Y

$$IoU(Y, \hat{Y}) = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|}. \quad (2)$$

Then, if the IoU is greater than α, the predicted instance is considered a true detection and a false one otherwise.

While the exclusive use of IoU in a coarse localization problem such as detection is clear, several limitations arise from its use on the finer localization obtained by instance segmentation. For example, in Fig. 1, all the masks G_i have the same IoU with the ground truth even if they are relatively different (shifted mask, over-infra segmentation, coarse mask). They would be equally viable as a true prediction in $mAP_{0.5}$ for example. Though, we clearly see that IoU fails to assess both contour detection errors (e.g. coarse mask in G_4) and content errors (e.g. mask G_3 obtained by enlarging GT).

The main limit of the mAP based on IoU in instance segmentation problems is that it only relies on pixel inclusion in a mask. This is a known issue and authors such as Csurka *et al.* [6] recommend the use of multiple metrics to capture different aspects of semantic segmentation. In [6], IoU and a contour discrepancy metric are for instance considered. This paper builds upon previous works and introduces in the next section an evaluation protocol for instance segmentation in the fashion context.

3. Multi level evaluation

In order to deeply analyze the impact of an inaccurate predicted mask, we propose in this section a multi level evaluation in the context of clothes instance segmentation in the fashion industry. The proposed multi level evaluation exploits all the information available at the pixel level: mask membership (global), location (contour) and color (content). In the following subsections, we introduce different metrics and select one candidate metric for each level of evaluation. Their application to segmentation instances is illustrated in Fig. 1.

3.1. Global level

Segmentation is a pixel classification task that predicts a class label for each pixel. Hence, standard evaluation protocols focus on the evaluation of the accuracy of the mask region estimated during the classification. Clustering evaluation (*e.g.* rand index) or information theoretic based metrics (*e.g.* mutual information) can be used to evaluate pixel classification. In this work, we consider overlap coefficients, which are popular metrics for segmentation. Overlap coefficients include the Dice index and the IoU.

We propose to rely on IoU, defined in (2), as an efficient reading of global mask quality. Well established in the literature, it benefits an ease of use for comparison purposes. Moreover, the IoU is independent from pixels content and localization and evaluates the segmentation in itself.

3.2. Contour level

Segmentation can be formulated as a contour detection task, where the boundary of a mask is a closed contour to be detected. Evaluation of segmentation can thus be done with contour discrepancy metrics, in order to have an information on the accuracy of the boundary of segmented objects. The Hausdorff distance (HD) and the boundary f1-measure (BF) [13, 6] are examples of such metrics. These approaches are nevertheless either too expensive to compute (HD) or difficult to analyze (lack of expressiveness for BF).

Hence we propose to use the boundary Jaccard [7], which improves boundary f1-measure [13, 6]. Boundary Jaccard (BJ) compares the contours $\partial\hat{Y}$ of the predicted mask $\hat{Y}$ with the ground truth ones ∂Y. To express BJ, we define d as the distance from a pixel x to a mask B: $d(x, B) = \inf_{y \in B} ||x - y||$ and the metric D between a contour ∂A and a mask B, for an accuracy threshold $\theta > 0$

$$D(\partial A, B) = \sum_{x \in \partial A,\ d(x,B) < \theta} \left(1 - (d(x, B)/\theta)^2\right), \quad (3)$$

We finally obtain the Boundary Jaccard as

$$BJ(Y, \hat{Y}) = \frac{D(\partial\hat{Y}, Y) + D(\partial Y, \hat{Y})}{|Y| + |\hat{Y}|}. \quad (4)$$

Notice that all contour pixels that are at a distance greater than θ to the mask have a zero contribution to the BJ index. For the threshold θ, the authors of [7] proposed a value of 0.75% of the image diagonal. In our experiment we set θ to a fixed value of 5, according to the images size.

For fashion related tasks it is crucial that the masks preserve clothes shape and localization. BJ gives insight on the model ability to match ground truth contours (see Fig. 1).

3.3. Content level

In order to evaluate the quality of the segmented content, we now propose to analyze the color distribution within the segmented masks. In practice, we consider the color distribution of the pixels in a mask as a discrete $3D$ histogram of n bins defined in the L*a*b* CIE76 color space [21]. The evaluation then consists of a comparison between color histograms of ground truth and estimated masks. Discrete histogram comparison tools can be divided into four main categories [18]: heuristic histogram distance, statistical test, information theoretic divergence and ground distances.

We here propose to rely on the Earth Mover's Distance (EMD) [22]. This ground distance has indeed proven to be a robust metric for image retrieval [16], color transfer [17] or image segmentation [15]. Contrary to classical bin-to-bin measures such as Kullback-Leibler divergence, EMD is naturally designed to take into account empty bins. On the other hand, there exists no explicit formula to compute EMD between histograms defined on spaces of dimension greater than one. As we deal with $3D$ histograms, we need to solve a linear optimization problem to compute EMD.

We denote as $h_{\hat{Y}}$ and h_Y the color histograms of the pixels respectively contained in estimated and ground truth masks $\hat{Y}$ and Y. EMD is obtained from a flow f that gives the minimal cost for transporting $h_{\hat{Y}}$ to h_Y, given a $n \times n$ matrix which components $c_{i,j}$ represents a cost between bins i and j. The optimal value f_{ij}, that indicates the portion of mass in the histogram bin $h_{\hat{Y}}(i)$ transported to the histogram bin $h_Y(j)$, is estimated by solving

$$EMD(h_Y, h_{\hat{Y}}) = \inf_f \sum_i^n \sum_j^n f_{i,j} c_{i,j} \tag{5}$$

subject to the constraints: (i) $\sum_i^n f_{i,j} = h_{\hat{Y}}(j)$, $j : 1 \cdots n$, (ii) $\sum_j^n f_{i,j} = h_Y(i)$, $j : 1 \cdots n$ and (iii) $f_{i,j} \geq 0$, $i, j = 1 \cdots n$. We use as cost matrix $c_{i,j} = ||b_i - b_j||$, with $\{b_i\}_{i=1}^n$ the centers of histogram bins, and we solve (5) with a linear solver. In order to define a similarity from EMD, we propose the following nonlinear transform:

$$sEMD(h_Y, h_{\hat{Y}}) = e^{(-\beta . EMD(h_Y, h_{\hat{Y}}))}. \tag{6}$$

Numerical experiments suggest that taking $\beta = 5$ and 16-bins histograms is a relevant choice for discriminating acceptable color histograms $h_{\hat{Y}}$ from those that are visually too different from h_Y.

With this content evaluation, we are able to estimate the color accuracy of the estimated masks. For certain fashion applications, extracting clothes fabrics can be as important as the clothes themselves. Being able to quantify errors based on over-segmentation (*e.g.* including background, other garment, *etc.*) and under-segmentation (*e.g.* missing clothing parts made up with different fabrics) is extremely valuable (see *e.g.* $sEMD$ in Figure 1).

3.4. Corpus evaluation

The three previous levels of evaluation concern individual masks. In order to realize an analysis of the results on a whole corpus, the information has to be aggregated.

First, we evaluate the distribution of values for each metric by selecting their means $m_{IoU}, m_{BJ}, m_{sEMD}$. Second, we propose to use the mean average precision (mAP) presented in Section 2 where IoU, BJ, $sEMD$ will be employed as the underlying true positive discrimination metrics with an associated α threshold. Each are named respectively: $mAP_{IoU\alpha}, mAP_{BJ\alpha}, mAP_{sEMD\alpha}$.

4. Experiment

The dataset Deepfashion2 [8] was assembled to answer multiple fashion related tasks: detection, landmark detection, pose estimation, segmentation, product retrieval. Deepfashion2 is currently the largest dataset containing mask annotations. The dataset was originally presented with the results of Mask R-CNN [10] method and a segmentation evaluation in terms of mAP_{IoU}.

For our experiments we consider the same data splitting as in [8] for training and evaluation (*i.e.* 52,490 instances in 32,153 images isolated for evaluation). The machine used to conduct the experiments was equipped with a Tesla P100 GPU. As the evaluated methods require non-negligible computation time for training, the final evaluation is realized after 5 epochs.

4.1. Evaluated methods

There exist two main categories of detection methods using CNN. The first one, popularized by Faster R-CNN [20], have two steps: a region proposal step and a region classification-refining step. The second category contains single shot approaches (*e.g.* YOLO [19]). Fixed grids are considered to reduce the complexity of the region proposal step, but theoretically at the expanse of prediction quality.

When it was proposed, the two steps method Mask R-CNN [10] achieved state-of-the-art results, by adding a segmentation step to the bounding box predicted by Faster R-CNN architecture. Many improvements have then been introduced concerning different aspects of the pipeline. MS R-CNN [11] focuses on improving the mask score and training loss. In particular it adds a prediction by regression of

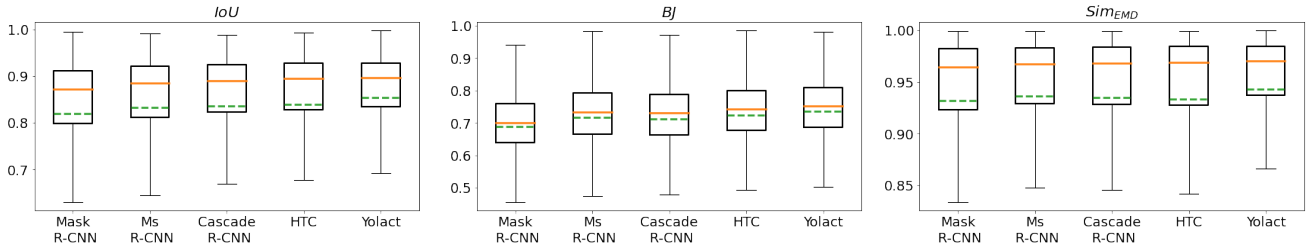

Figure 2: Box plot of the applied metrics, on the paired ground truth and first prediction ordered by score, distributions without outliers, green dashed line is the distribution mean, the yellow line is median.

Architecture	m_{IoU}	mAP_{IoU}			m_{BJ}	mAP_{BJ}			m_{sEMD}	mAP_{sEMD}		
		$[0.5, 0.95]$	0.5	0.75		$[0.5, 0.95]$	0.5	0.75		$[0.5, 0.95]$	0.5	0.75
Mask R-CNN	0.820	0.399	0.567	0.464	0.690	0.226	0.568	0.081	0.932	0.522	0.580	0.560
Mask scoring R-CNN	0.832	0.421	0.567	0.490	0.717	0.264	0.569	0.174	0.936	0.530	0.577	0.563
Cascade Mask R-CNN	0.836	0.424	0.577	0.493	0.713	0.257	0.578	0.145	0.935	0.533	0.589	0.568
Hybrid Task Cascade	0.838	0.440	0.594	0.508	0.725	0.283	0.600	0.187	0.934	0.547	0.608	0.584
Yolact	**0.854**	**0.516**	**0.687**	**0.601**	**0.737**	**0.341**	**0.689**	**0.265**	**0.943**	**0.642**	**0.699**	**0.679**

Table 1: Evaluation after 5 training epochs, m_{IoU}, m_{BJ} and m_{sEMD} are computed with the distributions presented in Figure 2, regarding the mAP the second line correspond to the α's interval, values

the IoU to the predicted mask score. Cascade R-CNN [4] is a multi-stage detection architecture composed of multiple detector. Each detectors are trained sequentially with increasing rigorousness in predictions. Hybrid Task Cascade (HTC) [5] improves on Cascade R-CNN by putting forward a intertwined detection and segmentation chain instead of the two task being cascaded separately and adding a context branch in the architecture.

Yolact [3] is a single shot methodology for real-time instance segmentation. The architecture mimics Mask R-CNN approach but on the single shot detector YOLO [19]. The mask predicted results from a linear combination of mask prototypes and instance coefficient generated in two different branches. Boyla *et al*. [3] claim that this approach could be adapted to almost any detection architectures.

To sum up, we trained 5 different segmentation architectures for the evaluation : Mask R-CNN [10], Mask Scoring R-CNN (MS R-CNN) [11], Cascade Mask R-CNN [4], Hybrid Task Cascade (HTC) [5], Yolact [3]. Note that all these architectures are built on top of a ResNet-50 backbone.

4.2. Analysis of results

As reported on Table 1, Yolact gives the best results after 5 training epochs. Moreover, its training time is almost three time faster than Mask R-CNN (Table 2). HTC, is the runner-up in term of mask quality and the best of two steps approaches. However, the architecture complexity increases training time by approximately 50% compared to Mask R-CNN and a factor four with Yolact. MS R-CNN has the second best m_{EMD} (Table 1). Its worst predictions are better than the ones of others two steps methods (see boxplot lower bounds in Fig. 2). It can also be noticed that MS R-CNN is better than Cascade R-CNN in terms of contour accuracy, even if this latter model performs better for the global level evaluation. The $mAP_{Iou\alpha}$, the $mAP_{BJ\alpha}$, and the $mAP_{sEMD\alpha}$, decreases slightly faster with increasing α, for both Mask R-CNN and Cascade R-CNN when compared to other models. This suggests that the two methods produce much more false predictions.

5. Conclusion and future work

In this paper, we propose a three levels evaluation framework for instance segmentation. We applied our methodology to clothes segmentation obtained from five state-of-the-art segmentation models. The framework proves to give useful insights on models inference, adding interpretability to results. We show that Yolact obtains the best performances after an early stopping of 5 epochs. Notice that the evaluation trends were reinforced during training and we postulate that more epochs will only accentuate the current gaps between the five models. We plan to evaluate other single shot instance segmentation methods in the future.

Finally, possible improvements of the evaluation framework include a better normalization for the $sEMD$, the reduction of EMD computational cost, and a texture evaluation for the content level.

	Mask R-CNN	MS R-CNN	Cascade R-CNN	HTC	Yolact
inference time *s/image*	0.11	0.11	0.13	0.22	0.07
training time *h/epoch*	10.94	11.20	14.68	16.87	3.89

Table 2: Time complexity of the 5 evaluated methods.

References

[1] M. Bastan, A. Ramisa, and M. Tek. Cross-modal fashion product search with transformer-based embeddings. In *CVPR Worksh.*, 2020. 1

[2] M. Bhattacharyya and S. Nag. Hybrid style siamese network: Incorporating style loss in complementary apparels retrieval. In *CVPR Worksh.*, 2020. 1

[3] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee. Yolact: Real-time instance segmentation. In *ICCV*, 2019. 1, 4

[4] Z. Cai and N. Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE Trans. PAMI*, 2019. 1, 4

[5] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 1, 4

[6] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan. What is a good evaluation measure for semantic segmentation? In *BMVC*, volume 27, 2013. 2

[7] E. Fernandez-Moral, R. Martins, D. Wolf, and P. Rives. A new metric for evaluating semantic segmentation: leveraging global and contour accuracy. In *IEEE intelligent vehicles symposium*, 2018. 2, 3

[8] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *CVPR*, 2019. 3

[9] S. Goree and D. Crandall. Studying empirical color harmony in design. In *CVPR Worksh.*, 2020. 1

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 3, 4

[11] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang. Mask scoring r-cnn. In *CVPR*, 2019. 1, 3, 4

[12] Y.-H. Ji, H. Jun, I. Kim, J. Kim, Y. Kim, B. Ko, H.-K. Kook, J. Lee, S. Lee, and S. Park. An effective pipeline for a real-world clothes retrieval system. In *CVPR Worksh.*, 2020. 1

[13] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. PAMI*, 26(5):530–549, 2004. 2

[14] M. Minar, T. Tuan, H. Ahn, P. Rosin, and Y. Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *CVPR Worksh.*, 2020. 1

[15] N. Papadakis and J. Rabin. Convex histogram-based joint image segmentation with regularized optimal transport cost. *Journ. of Math. Im. and Vis.*, 59(2):161–186, 2017. 3

[16] O. Pele and M. Werman. Fast and robust earth mover's distances. In *ICCV*, 2009. 3

[17] F. Pitié, A. C. Kokaram, and R. Dahyot. Automated colour grading using colour distribution transfer. *Comp. Vis. and Im. Underst.*, 107(1-2):123–137, 2007. 3

[18] J. Puzicha, J. M. Buhmann, Y. Rubner, and C. Tomasi. Empirical evaluation of dissimilarity measures for color and texture. In *ICCV*, 1999. 3

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 3, 4

[20] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inform. Process. Syst.*, 2015. 3

[21] A. Robertson, R. Lozano, D. Alman, S. Orchard, J. Keitch, R. Connely, L. Graham, W. Acree, R. John, R. Hoban, et al. Cie recommendations on uniform color spaces, color-difference equations, and metric color terms. *Color Res. Appl*, 2:5–6, 1977. 3

[22] Y. Rubner, C. Tomasi, and L. Guibas. A metric for distributions with applications to image databases. In *ICCV*, 1998. 3