Body Segmentation Using Multi-task Learning

Julijan Jug[1,†], Ajda Lampe[1,2,†], Vitomir Štruc[2], Peter Peer[1]

[1]*Faculty of Computer and Information Science,* [2]*Faculty of Electrical Engineering*
University of Ljubljana, SI-1000 Ljubljana, Slovenia
julijan.jug@gmail.com, {ajda.lampe, vitomir.struc}@fe.uni-lj.si, peter.peer@fri.uni-lj.si

Abstract—**Body segmentation is an important step in many computer vision problems involving human images and one of the key components that affects the performance of all downstream tasks. Several prior works have approached this problem using a multi-task model that exploits correlations between different tasks to improve segmentation performance. Based on the success of such solutions, we present in this paper a novel multi-task model for human segmentation/parsing that involves three tasks, i.e., (i) keypoint-based skeleton estimation, (ii) dense pose prediction, and (iii) human-body segmentation. The main idea behind the proposed Segmentation–Pose–DensePose model (or SPD for short) is to learn a better segmentation model by sharing knowledge across different, yet related tasks. SPD is based on a shared deep neural network backbone that branches off into three task-specific model heads and is learned using a multi-task optimization objective. The performance of the model is analysed through rigorous experiments on the LIP and ATR datasets and in comparison to a recent (state-of-the-art) multi-task body-segmentation model. Comprehensive ablation studies are also presented. Our experimental results show that the proposed multi-task (segmentation) model is highly competitive and that the introduction of additional tasks contributes towards a higher overall segmentation performance.**

Index Terms—**computer vision, segmentation, human body parsing, multi-task learning**

I. INTRODUCTION

In recent years, great progress has been made in the field of computer vision. Modern generative models, such as GANs, have made it possible to generate photorealistic images with convincing visual quality. Much research is also focused on the application of such models. One such challenge is the generation of photorealistic images of people in desired clothing or the problem of virtual try-on [1], [2]. Such applications have great potential for use in online clothing stores and enhance the user experience of online shopping. With the development of deep neural networks, there has also been a great leap in the field of semantic segmentation [3], [4]. However, there is still much room for improvement in certain areas, such as human body segmentation. Currently, the best segmentation models still do not perform as well as they should for applications such as virtual clothing try-on. Most of the problems with current models are caused by images taken under less than ideal conditions and partially obscured views of the subject.

A significant amount of research has been conducted recently to improve such models by using additional information

Fig. 1. This example shows that the pose and dense pose subtasks provide helpful contextual and structural information about the human body. The second image shows the segmentation mask produced by our multi-task model containing segmentation and pose estimation tasks. The third image shows the segmentation mask created by our multi-task model with the additional task of dense pose estimation. We can see that the additional task of dense pose estimation significantly improves the segmentation performance.

to drive and support segmentation models. By providing additional contextual information, it is assumed that the model may obtain a better understanding of the image content and human anatomy. Existing work is, therefore, looking at combining segmentation models with other related tasks in so-called multi-task architectures. Most commonly existing models include pose estimation as a supporting task, e.g., [5]. Previous research has also shown that using a multi-task learning contributes to the quality of human segmentation. Based on this insight, we explore in this paper additional possibilities for extending this type of models with additional tasks that could further aid the segmentation process.

While most existing work includes pose estimation as a supporting task, our work focuses on improving the quality of segmentation results by utilizing an additional task. To this end, we propose a new architecture of a multi-task model that includes the task of inference of skeletal position or posture and dense pose in addition to the task of body segmentation. To this end, we propose a novel multi-task segmentation model called SPD, which considers all three tasks. The letters in the name represent each task: S – segmentation, P – pose, and D – dense pose. We propose a multi-task architecture based on a shared backbone neural network using three specialized branches on top, one for each of the selected tasks. The purpose of such an approach is to improve the segmentation task. We evaluate the proposed model on the LIP and ATR dataset and report highly encouraging results. We also perform extensive ablation studies to support our hypothesis that adding tasks improves the overall performance of the model.

The main contributions of this paper are:

- We present SPD, a novel multi-task model for human body segmentation that includes pose estimation and

dense pose prediction tasks.
- We show that adding additional tasks improves performance for the primary task.

II. RELATED WORK

One of the more specialized application domains of semantic segmentation is the segmentation of the human body and clothing. The need for such segmentation algorithms arises from the requirements of various vision systems related to human image analysis, such as virtual clothing try-on [1], [2] or re-identification [6]. Recently, much research has been done on human segmentation [7]–[9] using deep convolutional neural networks. The disadvantage of these models is that they do not take into account the structure or anatomy of the human body, so the segmentations often contain errors that are unreasonable from a human perspective.A considerable amount of research has, therefore, focused on solving this problem by incorporating additional information into the segmentation procedure related to body posture and anatomy.

One way to introduce supporting information to the model is the multi-task learning approach, where the model is simultaneously trained to solve multiple tasks. Due to the good results in recent years, multi-purpose learning has been widely used in various natural language processing and computer vision applications [10]–[12]. Gong *et al.* [5], for example, proposed a model that predicts semantic segmentation masks and estimates body joint positions based on the generated segmentation map. The model is optimized based on the quality of both the segmentation map and the joint locations to ensure it learns a semantically consistent representation of the human body. Liang *et al.* [13] build on this approach by using a common base network, followed by two smaller modules, specialized for joints estimation and semantic segmentation. The modules are built in a two-stage coarse-to-fine manner and share the intermediate coarse results. The proposed model, called JPPNet, achieves impressive results and outperforms previous work in a convincing manner. However, there is still room for improvement in the model's body representation, as some structural flaws persist. Given the promising results, we explore the potential of introducing an additional task in improving the final semantic segmentation result.

III. METHODOLOGY

We propose a multi-task model, called SPD, for human body parsing that is learned based on three distinct tasks: segmentation mask generation, keypoint-based pose estimation, and dense pose prediction [14]. The model is inspired by the success of existing multi-task models, such as JPPNet [13], that have been shown to ensure competitive performance, while also exhibiting desirable architectural features.

A. Model Overview

Fig. 2 shows the basic architecture of our model, which consists of a backbone feature extractor and three distinct branches: (i) one for human body segmentation, (ii) one for key-point based pose estimation, and (iii) one dense pose
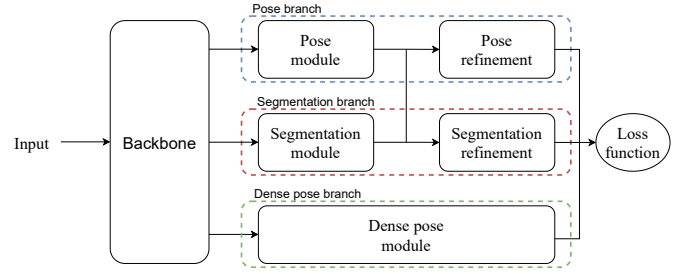

Fig. 2. High-level architectural diagram of the proposed SPD model. The common ResNet backbone of the SPD model is shared between three specialized model branches designed specifically for human body segmentation, skeleton/pose prediction, and dense pose estimation.

prediction. The main goal of the model is to ensure efficient body part segmentation, so the segmentation branch is treated as the main component of the model, whereas the remaining two branches perform auxiliary tasks. The main backbone model common to all tasks is a ResNet-101 [15] deep residual network, which consists of 101 convolutional layers arranged across 5 residual blocks. In the SPD model, part of this backbone is shared between the three branches, which acts as a link between the three considered tasks.

The three branches allow for the definition of three separate learning objectives, i.e., one per tasks, that are then used jointly to learn the model. Specifically, the overall loss function used with SPD is calculated as a weighted sum of the three task-specific losses, i.e.:

$$\mathcal{L} = \lambda_s \mathcal{L}_s + \lambda_p \mathcal{L}_p + \lambda_d \mathcal{L}_d, \tag{1}$$

where λ_s, λ_p and λ_d are balancing weights corresponding to the segmentation loss $\mathcal{L}_s$, the keypoint-based pose loss $\mathcal{L}_p$, and the dense-pose loss $\mathcal{L}_d$, respectively . Empirically, we chose a higher weight for the segmentation part of the loss function and lower values for the other two tasks to ensure that the segmetnation task is given preference in the optimization procedure. We set $\lambda_s = 1$, $\lambda_p = 0.8$ and $\lambda_d = 0.6$ based on preliminary experiments to provide a good trade-off between the three tasks. The individual loss functions are presented in the following subsections.

B. Segmentation Branch

Usually, only information from the ground truth segmentation masks is used to learn the task of segmenting individuals. In our approach, we also incorporate contextual information of the skeleton directly into the segmentation network. Fig. 3 shows a high-level overview of the components in the segmentation branch. As can be seen, the output of the fifth residual block is used as the initial representation for the segmentation barnch. To generate the an initial estimate of the segmentation mask, an additional layer of Atrous Spatial Pyramid Pooling (ASPP) is used on top of the extracted ResNet features. ASPP performs multiple convolutions over the input data at different sampling rates and mask sizes, capturing objects and contextual information at different scales. In parallel to the ASPP component, we create what we call *segmentation context*
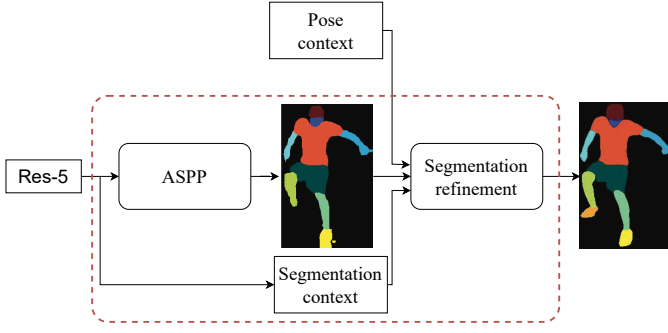
Fig. 3. Overview of the segmentation branch of the SPD model. The branch consists of two parts, where the first generates an initial segmentation result based on features produced by the backbone model, whereas the second refines this initial estimate using different types of input information - also from other branches.
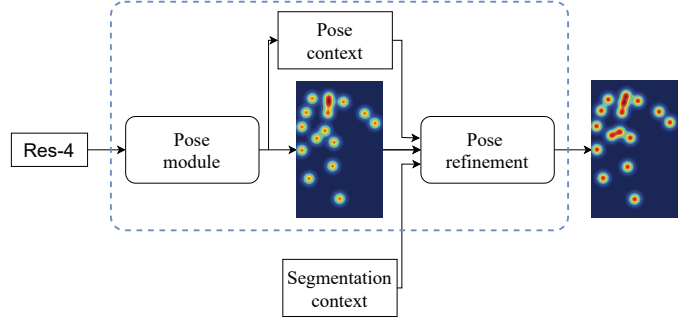

Fig. 4. Overview of the pose estimation branch of the SPD model. The branch consists of two parts, where the first generates an initial keypoint prediction based on features produced by the backbone model, whereas the second refines this initial estimate using different types of input information - also from other branches.

by processing the output of the fifth residual layer through two additional convolutional layers. This context is later used in the second stage of the segmentation branch along with other sources of information to further refine the segmentation results.

The refinement network in the second part of the segmentation branch takes the segmentation context as well as the initial (rough) estimate of the segmentation masks as input and combines these inputs with what we call *pose context* - a representation generated by the pose estimation branch of the model. This is followed by four levels of convolutions with the purpose of capturing the local context and learning the key connections between the pose and segmentation contexts. The result of these convolutional layers is reshaped and passed through another ASPP component. This last ASPP component, thus, generates the final segmentation masks. The segmentation loss defined on top of this branch is expressed in terms of pixel-wise softmax cross entropy, i.e.:

$$\mathcal{L}_s = \frac{1}{M} \sum_{k=1}^{K} \sum_{m=1}^{M} y_m^k \times log(h_\theta(x_m, k)), \quad (2)$$

where M is the number of samples, K is the number of segmentation classes, y_m^k is the target classification for a sample m and a class k. The input sample is denoted by x and the prediction model by h.

C. Pose Estimation Branch

Fig. 4 shows a high-level overview of the components involved in creating pose representations, i.e., keypoint of the human skeleton. Unlike in the segmentation branch, the input to the pose branch is the output of the fourth residual block, following the suggestions from [13]. The initial pose module in this branch consists of 8 convolutional layers, the initial six obtain skeleton features, and the two on top produce the first version of the skeleton representation in the form of a tensor with sixteen coordinates of skeleton joints. Similarly as in the segmentation branch, a refinement step is used in the second stage of this branch that take the initial pose predictions, pose context and the segmentation context, produced by the

segmentation branch as input, and then applies 4 levels of convolutions over these input to capture the local context at different scales. Finally, two additional convolutional layers are utilized to generate the refined skeleton keypoints. An L2 loss is defined on top of the branch to facilitate training, i.e.:

$$\mathcal{L}_p = \frac{1}{2N} \sum_{i=1}^{N} \|p_i - p_i{}'\|^2, \quad (3)$$

where N represents the number of defined joints in the skeleton, $p_i{}'$ the predicted coordinates of the joint, and p_i the annotated coordinates of the joint.

D. Dense Pose Branch

Fig. 5 presents the architecture of the dense pose branch. Similarly as in the pose branch, we use the first the output of the fourth residual block of the backbone model for the initial encoding. Following the ResNet network is a module for sampling regions of interest (ROIs), which is used to (cascadely) capture local contexts at various scales. Attached to the ROI pooling module is a *dense pose* head, composed of two dedicated CNN heads, a classification head and a regression head. The first head is used to assign the image elements to the corresponding body segment, i.e., the classification of component I. The second head determines the position of the image elements within the corresponding segments, i.e., it is used to determine the components U and V.

The loss function for the dense pose branch consists of two parts. The first part relates to the component I and is computed in the same way as in the main segmentation task, i.e., using cross entropy. The second part, which refers to the coordinates U and V, is computed using the Huber loss function:

$$\mathcal{L}_d = \sum_{m=1}^{M} CSE(x_I) \cdot L_1(x_U, x_V), \quad (4)$$

where x_I, x_U, x_V are the components of the depth representation, CSE is the transverse entropy function for the segmentation part and L_1 is the Huber loss function for the position part.
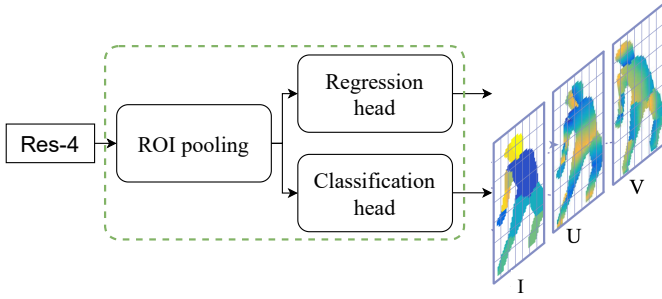
Fig. 5. The figure shows a high-level diagram of the architecture of the DensePose model. Dense pose components visualization is taken from a DensePose article [14].

Fig. 6. Example image from the LIP dataset and the corresponding segmentation masks and 16-point skeleton markup.

E. Training Details

The models are trained on an Nvidia 1080Ti GPU with 11 GB of memory. We empirically determined the weights $\lambda_s = 1$, $\lambda_p = 0.8$, $\lambda_d = 0.6$ for the objective in (1). The full model was trained for 400.000 iterations.

IV. EXPERIMENTS AND RESULTS

In this section, we present the datasets selected for the experiments. We describe the protocol used to evaluate the proposed SPD model and discuss the performance measures utilized for performance reporting for all three model tasks. We then comment on and analyze the results of the model. We also perform an ablation analysis to demonstrate the contribution of each task to the final accuracy of the proposed model. Finally, we present examples of the generated segmentation masks and analyze them qualitatively.

A. Datasets

Dataset selection plays an important part in the training of the proposed SPD model. For our purposes, we used several datasets containing images of people in different clothing, situations, contexts, and body positions. A particular challenge of our multi-task modeling approach is the need for a database that contains several different types of annotations.

For learning a multi-task model that includes the generation of segmentations, body poses, and dense pose representations, we need a dataset that contains all three types of annotations. To this end, we selected the LIP dataset [5] that contains segmentation and skeleton annotations for over 50.000 images. An example image with the segmentation masks and pose annotations from this dataset is presented in Fig. 6. For dense pose annotations, we used the COCO [16] database, which is a superset of LIP. We merged the annotations from both dataset to generate the reference data needed to train the SPD model. In the final setup we have dense-pose annotations for all input images, a 19-class markup for the segmentation task, and a 16-point markup for the skeleton keypoints. To evaluate the performance of all tasks of SPD, we use a hold-out set from LIP, as well as images from the ATR [17] dataset.

B. Performance Measures

Following standard evaluation methodology, we use four performance measures to report performance for the segmentation task, i.e., the Jaccard index IoU, precision, recall, and the $F1$ score [18], [19]. The first measure is the Jaccard index or the weighted average of the intersection over union. The measure IoU (intersection over union) is defined as:

$$IoU = \sum_{i=1}^{K} \frac{S_i{}' \cap S_i}{S_i{}' \cup S_i}, \tag{5}$$

where S' represents the predicted area and S represents the annotated area of the i-th instance class and K is the number of annotated reference classes. The maximum value of $IoU = 1$ indicates ideal performance. When looking at semantic segmentation as a pixel-level classification problem, precision (6) is defined as the ratio of correctly classified pixels among all pixels classified to a class, whereas recall (7) is the fraction of correctly classified pixels among all pixels belonging to a class, i.e.:

$$Precision = \frac{TP}{TP + FP}, \tag{6}$$

$$Recall = \frac{TP}{TP + FN}, \tag{7}$$

where TP, FP, TN and FN denote true positives, false positives, true negatives and false negatives, respectively. The $F1$ score is the harmonic mean between precision and recall:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{8}$$

For the pose estimation task we report the mean Euclidean distance (mED) between the predicted $p_i{}'$ and reference pose key points p_i. The measure is defined as:

$$mED = \sum_{i=1}^{N} d_{L_2}(p_i, p_i{}'), \tag{9}$$

where $d_{L_2}(\cdot)$ is the Euclidean distance function, and $N = 16$ is the overall number of annotated key points.

For the dense pose prediction task, we use a measure of geodesic point similarity between the generated and reference

Experiment	Model	Performance measures			
		$IoU \uparrow$	$Pr \uparrow$	$Rec \uparrow$	$F1 \uparrow$
Comparison	JPPNet [13]	0.538	0.68	0.66	0.66
	SPD (ours)	**0.547**	**0.76**	**0.68**	**0.71**
Ablation study	SP	0.535	0.74	0.52	0.63
	SD	0.478	0.67	0.50	0.57
	S	0.483	0.62	0.49	0.54

Experiment	Model	Performance measures			
		$IoU \uparrow$	$Pr \uparrow$	$Rec \uparrow$	$F1 \uparrow$
Comparison	JPPNet [13]	$0,464$	$0,66$	$0,67$	$0,66$
	SPD (ours)	$\mathbf{0,472}$	$0,67$	$\mathbf{0,70}$	$\mathbf{0,68}$
Ablation study	SP	$0,423$	$\mathbf{0,69}$	$0,53$	$0,60$
	SD	$0,340$	$0,59$	$0,44$	$0,50$
	S	$0,291$	$0,50$	$0,56$	$0,52$

dense-pose body representations, as defined in [14]. The measure is defined as follows:

$$GPS = \frac{1}{|P|} \sum_{p_i \in P} exp \left(\frac{-d(\hat{p}_i, p_i)^2}{2k(p_i)^2} \right). \quad (10)$$

In the above definition, P represents the set of annotated surface points, $| \cdot |$ is the set cardinality, $\hat{p}_i$ denotes the i-th predicted point on the surface, and p_i the corresponding annotated point on the person's surface. The function d represents the geodesic distance between the points and k the normalization factor specific to each body part. The values of the normalization factors are taken from [14].

C. Segmentation Results and Ablations

Comparison with the State-Of-The-Art. With the proposed SPD model we aim to improve on the results of existing body segmentation models. Specifically, we build on the recent JPPNet approach from [13] and, therefore, include this approach for baseline comparisons in the experiments. Table I shows the results for the segmentation task on the hold-out set of LIP. As can be seen, on the LIP dataset, the SPD model achieves an IoU result of 0.547, compared to the JPPNet model, which results in a score of 0.538. In terms of the $F1$ score, the proposed model outperforms JPPNet by approximately 5%. Similar performance improvements are also observed for precision and recall. To further verify the performance of SPD on an independent dataset with characteristics different from the training data, we also evaluate our model on the ATR dataset. The segmentation results in Table II again show that SPD outperforms JPPNet in terms of all reported performance metrics. We attribute the observed performance gains to the interaction of the three different tasks considered during training, which allow our model to better learn how to efficiently parse images of humans and generate reliable segmentation masks across a diverse set of image characteristics.

Ablation Study. To demonstrate the importance of all tasks in the multi-task design of SPD, we perform an ablation study, where different tasks are removed from the overall model. Three additional models are implemented and trained for this experiment, i.e.: (i) the SPD model without the dense-pose prediction task (SP hereafter), (ii) the SPD model without the keypoint-based pose prediction task (SD hereafter), and (iii) the SPD model without both pose-related tasks (S hereafter).

The results of this experiment are presented in Tables I and II for the LIP and ATR datasets, respectively. It can be seen, that each added task provides the model with new useful information to improve the segmentation results. Removing the dense pose estimation task results in a drop of the segmentation performance across all performance scores. The removal of the keypoint-based pose estimation task has an even bigger adverse effect on performance. If both task are ablated, we observe the most significant performance degradation suggesting that both pose-related tasks provide important information for further improving segmentation results. Interestingly, we see larger performance drops on the ATR dataset than on LIP. This is likely a result of the fact that the model was trained on part of the data in LIP, so auxilary tasks are more critical when the characteristics of the data change. In the cross-dataset experiment on the ATR dataset, the added information from the dense-pose and keypoint-based pose estimation tasks is needed to produce competitive segmentation performance with SPD.

D. Results of Auxiliary Tasks

Because SPD is trained in a multi-task manner, it also produces predictions of skeleton/pose keypoints and dense-pose representations of the input images. To better understand the behavior of the model, we report here results for the keypoint prediction and dense-pose estimation tasks on the test part of the LIP dataset.

Keypoint Prediction. For the first experiment we evaluate three models, the proposed SPD, the reference JPPNet and SPD model without the dense-pose prediction task, i.e., SP. On the LIP test data, the JPPNet model results in the lowest mED value of 51.2 pixels, followed by the SPD model with a value of 55.01 pixels. The weakest model in this experiment is the SP model with a mED value of 56.82 pixels. These results suggest that the addition of the dense-pose estimation task clearly improves performance for the keypoint prediction task. However, the final results are inferior to JPPNet, due to the fact that the segmentation tasks was given higher priority in the balancing of the loss term in Eq. (1).

Dense-pose Prediction. The third task performed within the SPD model is the prediction of the dense depth representation of the body. Because JPPNet does not generate dense-pose predictions, we only report results for the complete SPD model and the model without the keypoint-based pose prediction

task, i.e., SD. On the LIP test data, the SPD model achieves a GPS score of 48.2% and the SD model with a score of 50.1%. The results show that adding the keypoint-based pose prediction task does not help to improve dense-pose estimation performance. Both models result is very similar GPS scores. This observation suggests that even though segmentation can benefit from the additional tasks, the balancing weights used in our optimization objective do not ensure consistent performance improvements across all considered tasks. Nevertheless, if dense-pose prediction is treated as the primary optimization target, improved results can also be expected for this task due to the multi-task learning.

E. Qualitative analysis

In this section, we present and analyze qualitative results generated by the segmentation branch of the SPD model. Fig. 7 shows a comparison of the segmentation results generated by SPD and JPPNet together with the original input images the ground truth segmentation masks for two selected images from the LIP dataset.

Input image

Annotation

JPPNet

SPD
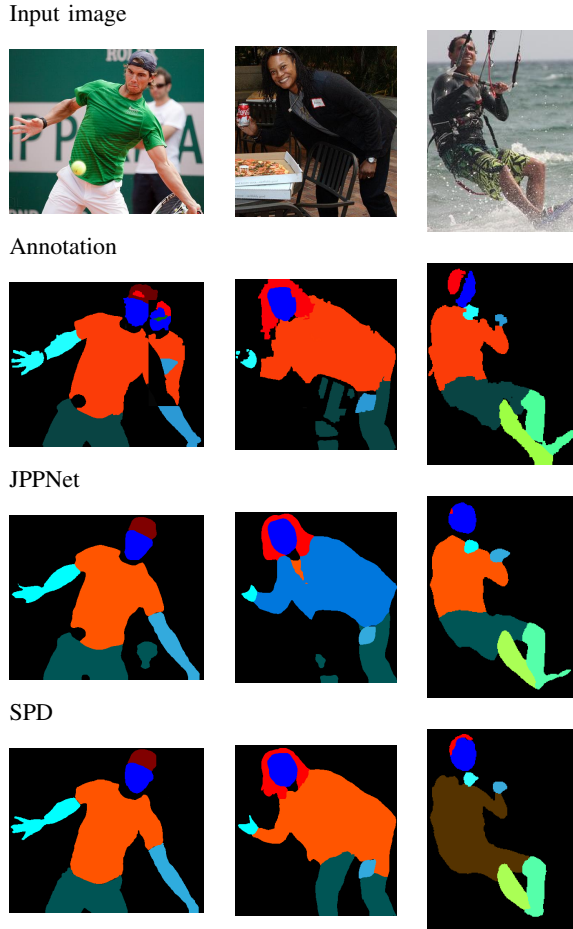
Fig. 7. Comparison of the segmentation results generated by the proposed SPD model and the competing JPPNet on selected images from the LIP dataset. The first row showd the selected input images, the second row shows the ground truth annotations, whereas the third and fourth row show results for JPPNet and SPD, respectively.

The first image shows a tennis player and a person in the background, who is out of focus and partially obscured. We can see that the SPD model is the only one that correctly detected only the player in the foreground. The competing model has problems with the person in the background, as it is very close to the tennis player in the foreground. The difference in segmentation quality is also visible in the definition of the fingers on the right hand, where the SPD model recognized individual fingers much better than the JPPNet modes. The second example image shows a woman partially hidden behind a chair. In this case, the JPPNet model omits the entire leg area, although it is still partially visible behind the chair. Despite the overlap, the SPD model recognizes the position of the leg and marks it correctly. Another unique feature of this image is the classification of the upper part of the garment. The upper part of the woman's body is annotated as an upper clothing class, JPPNet model falsely classifies it as a coat, while the SPD model correctly classifies the area as an upper clothing class, as a result of the contextual information provided by the other two tasks. In the third image, we see a man surfing on water. In this case, the JPPNet model results in the best segmentation according to the annotations, as it appropriately marks the upper part of the garment and separates that from the pants. Our model classifies the entire area as a one-piece jumpsuit, which is a reasonable classification given the appearance of the image from a human perspective.

V. CONCLUSION

In this work, we presented a multi-task segmentation model called SPD. In addition to the primary task of body segmentation, the model also includes the task of keypoint-based pose estimation and dense pose prediction. The segmentation part of the model was evaluated on the LIP and ATR datasets, and for both datasets SPD achieved better results than the reference model JPPNet. Furthermore, through rigorous ablation studies it was shown that models that considered a lesser number of tasks resulted in worse performance. In the ablation analysis, we presented the contribution of each of the tasks and found that using the skeleton and depth task together adds more value than using either one of them on its own. To further improve results, we plan to explore additional tasks in the learning procedure that could provide additional cues for the segmentation procedure.

REFERENCES

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "Viton: An image-based virtual try-on network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7543–7552.

[2] B. Fele, A. Lampe, P. Peer, and V. Struc, "C-vton: Context-driven image-based virtual try-on network," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," in *Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4. IEEE, Apr. 2018, pp. 834–848. [Online]. Available: https://doi.org/10.1109/tpami.2017.2699184

[4] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," vol. 43, no. 10. IEEE, Oct. 2021, pp. 3349–3364. [Online]. Available: https://doi.org/10.1109/tpami.2020.2983686

[5] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jul. 2017, pp. 556–567. [Online]. Available: https://doi.org/10.1109/cvpr.2017.715

[6] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2013, pp. 500–518. [Online]. Available: https://doi.org/10.1109/cvpr.2013.460

[7] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," in *Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 12. IEEE, Dec. 2015, pp. 2402–2414. [Online]. Available: https://doi.org/10.1109/tpami.2015.2408360

[8] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with local-global long short-term memory," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2016, pp. 710–724. [Online]. Available: https://doi.org/10.1109/cvpr.2016.347

[9] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," in *International Conference on Computer Vision (ICCV)*. IEEE, Dec. 2015, pp. 1150–1168. [Online]. Available: https://doi.org/10.1109/iccv.2015.163

[10] I. Kokkinos, "UberNet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jul. 2017, pp. 1380–1410. [Online]. Available: https://doi.org/10.1109/cvpr.2017.579

[11] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *International Conference on Computer Vision (ICCV)*. IEEE, Dec. 2015, pp. 800–820. [Online]. Available: https://doi.org/10.1109/iccv.2015.304

[12] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *International Conference on Image Processing (ICIP)*. IEEE, Sep. 2019, pp. 630–647. [Online]. Available: https://doi.org/10.1109/icip.2019.8803050

[13] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," in *Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4. IEEE, Apr. 2019, pp. 871–885. [Online]. Available: https://doi.org/10.1109/tpami.2018.2820063

[14] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2018, pp. 1120–1135. [Online]. Available: https://doi.org/10.1109/cvpr.2018.00762

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision – ECCV*. Springer International Publishing, 2014, pp. 740–755. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1_48

[17] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," vol. 37, no. 12. IEEE, Dec. 2015, pp. 2402–2414. [Online]. Available: https://doi.org/10.1109/tpami.2015.2408360

[18] P. Rot, M. Vitek, K. Grm, Ž. Emeršič, P. Peer, and V. Štruc, "Deep sclera segmentation and recognition," in *Handbook of vascular biometrics*. Springer, Cham, 2020, pp. 395–432.

[19] Ž. Emeršič, D. Sušanj, B. Meden, P. Peer, and V. Štruc, "Contexednet: Context–aware ear detection in unconstrained settings," *IEEE Access*, vol. 9, pp. 145 175–145 190, 2021.

In the main part of the paper, we mostly focused on the presentation of segmentation results, since it is the main task that our model is addressing. Here, we present some additional visual results for segmentation, as well as for the supporting tasks.

A. Segmentation results on ATR

Fig. 8 shows qualitative segmentation results for the ATR dataset. Both JPPNet and SPD achieve similar performance on the first and third example image. It appears that both models have difficulty distinguishing between various types of upper clothing. The second image shows a non-typical example that is composed of two separate images. Our SPD model achieves far superior performance on the upper part of the image, whereas both models struggle to handle the lower part. This is most likely due to the models being trained to segment images of a single person.

B. Pose results

Figs. 9 and 10 show visual results of our pose module compared to the JPPNet on LIP and MPII datasets, respectively. In Fig. 9, the first column shows a simple example - a full image of a person where all limbs are well visible. The other two columns show more challenging examples, where the target is not fully visible in the image. Both models still provide good estimation of the upper body and limbs, but struggle with the legs due to occlusion or not being fully included in the image. In the last image, our model provides a better estimate of the occluded right leg than JPPNet. Both models achieve comparable results on the images from the MPII dataset which was not used for training any of the models. It again shows that both of the models have difficulty handling lower limb occlusions.

C. Dense Pose Results

Due to the dense pose task not being included in the JPPNet model, we compare our model's performance to the original DensePose model [14]. Visual results are shown in Figure 11. It is difficult for a human to visually evaluate the dense pose performance in detail, due to the nature of representation, however we observe that our model occasionally fails to cover all of the body area. Despite that, as pointed out in the main paper, the imperfect dense pose branch still directs the model learning enough to improve the overall segmentation predictions.

Input image
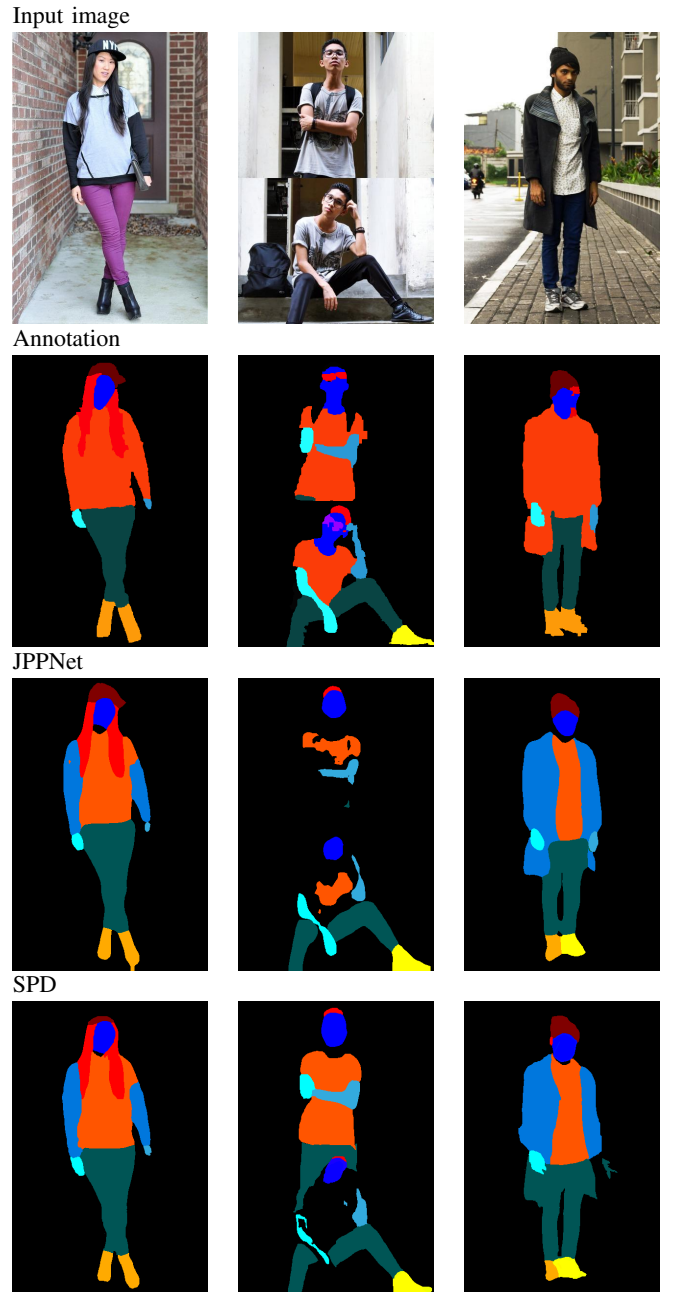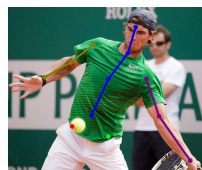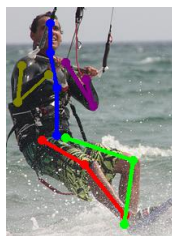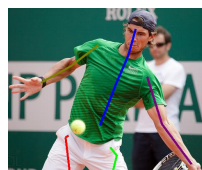
Annotation

JPPNet

SPD

Fig. 8. Comparison of segmentation masks for selected images from the database ATR.

Annotation

JPPNet
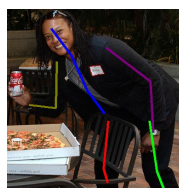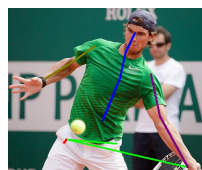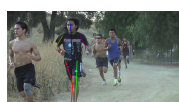

SPD
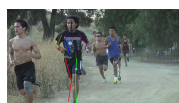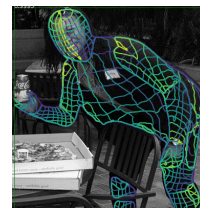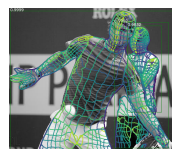

Fig. 9. Comparison of pose estimation results for selected images from the database LIP.
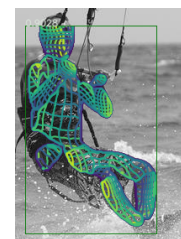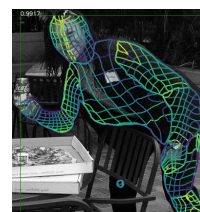
DensePose

SPD

Fig. 11. Comparison of dense pose estimation results for selected images from the database COCO.

Annotation

JPPNet

SPD

Fig. 10. Comparison of pose estimation results for selected images from the database MPII.