DESCRIBE ME IF YOU CAN! CHARACTERIZED INSTANCE-LEVEL HUMAN PARSING

Angelique Loesch[⋆,†] *Romaric Audigier*[⋆,†]

[⋆] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[†]Vision Lab, ThereSIS, Thales SIX GTS, Campus Polytechnique, Palaiseau, France
{angelique.loesch, romaric.audigier}@cea.fr

ABSTRACT

Several computer vision applications such as person search or online fashion rely on human description. The use of instance-level human parsing (HP) is therefore relevant since it localizes semantic attributes and body parts within a person. But how to characterize these attributes? To our knowledge, only some single-HP datasets describe attributes with some color, size and/or pattern characteristics. There is a lack of dataset for multi-HP in the wild with such characteristics. In this article, we propose the dataset CCIHP based on the multi-HP dataset CIHP, with 20 new labels covering these 3 kinds of characteristics.[1] In addition, we propose HPTR, a new bottom-up multi-task method based on transformers as a fast and scalable baseline. It is the fastest method of multi-HP state of the art while having precision comparable to the most precise bottom-up method. We hope this will encourage research for fast and accurate methods of precise human descriptions.

Index Terms— Human parsing, Characterized attributes, Dataset, Bottom-up segmentation, Scalability.

1. INTRODUCTION

Human semantic description is of utmost importance in many computer vision applications. It consists in automatically extracting semantic attributes corresponding to each person of an image. Attribute extraction is useful for many types of tasks, as *image content description*, *image generation* for virtual reality applications or *person retrieval* from a natural-description query, for security applications. Semantic attributes can also help visual signatures used in person *re-identification* and *person search* pipelines [1, 2].

Unlike *attribute classification* that aims to predict multiple tags to the image of a person [2] (sometimes deceived by nearby people/elements), *human parsing (HP)* [3, 4, 5, 6] aims to segment visible body parts, clothing and accessories at the pixel level. Localizing semantic attributes has several advantages. It provides a precise delineation of attributes necessary in augmented/virtual reality applications (entertainment, clothing retail...) [7]. It is more explainable than global tags (thus, more acceptable by human operators) and can cope with multiple-person descriptions by directly assigning localized people with attributes.



Fig. 1: Example of CCIHP ground truths (1st and 3rd rows) and our model predictions (2nd and 4th rows). First two rows: RGB image and legends, human instance, and semantic attribute maps. Last two rows: size, pattern and color maps.

Public datasets have been proposed for HP (cf. details in [8]). Many of them target retail/fashion applications and, thus, gather single-person images in controlled environments [9, 10, 11, 12, 13, 14]. Others present multiple-person images taken under in-the-wild environments [15, 16, 17, 6]. However, detecting the presence of attributes (e.g. pants, coat) is, in general, not sufficient to describe a person distinctively. Attributes should be characterized to provide a more complete and useful description. But no available dataset provide localized attributes with characteristics (such as color, size and pattern) in multi-person images. Indeed, even if some fashion-aimed datasets [11, 12, 13, 14] provide some color and pattern tags, they target single-person images with good resolution and pose. MHP v2.0 [17] provides many fine-grained attributes which partially include the size characteristic (e.g. pants vs. shorts, boot vs. shoe) but no color or pattern characteristics. This encouraged us to create a new

[1]CCIHP is available on https://kalisteo.cea.fr/index.php/free-resources/

HP dataset with multi-person characterized attributes. Thus, we have worked on CIHP [6], the largest existing multi-HP dataset, and annotated characteristics.

In order to cope with characterized HP, we also propose a method to serve as a baseline. Among existing HP methods, *single-person* methods make the assumption of a single person in the image, which means they do semantic segmentation and do not manage the attribute-to-person assignment problem [3, 18]. In contrast, *multi-person* methods cope with this problem by doing instance-level HP. They can be divided into three main categories. *Top-down two-stage* methods require human-instance segmentation of the image as an additional input [19, 20, 21, 4]. Their computation time highly depends on the number of people in the image. *Top-down one-stage* methods [22, 23, 5] predict both human instances and attributes. But computation time still depends of the number of people in the image because the top-down strategy forces to forward in the local parsing branch(es) as many times as the number of ROI candidates (people). *Bottom-up* methods also predict both human instances and attributes. Yet, their computation time does not depend on the number of people in the image [16, 17, 6]. However, these methods generally rely on expensive test-time augmentations, post-processing [16, 6] or heavy GAN architectures [17]. *Low and constant computation time* is essential when applying multi-HP on large amounts of possibly crowded images or videos. Thus, we propose a fast end-to-end bottom-up model based on transformers, which manages also the new task of attribute characterization, with low and constant computation time.

The contributions of this article are two-fold: (1) a new dataset for multi-HP in the wild with characteristics of localized attributes; (2) a bottom-up multi-task model for instance-level HP with characterized attributes. The proposed model does not need post-processing. It has low constant processing time, whatever the number of people per image, which makes it scalable and deployable. We hope this new dataset and baseline will encourage research for fast and accurate methods for more complete human descriptions.

2. PROPOSED DATASET

Our new dataset CCIHP (*Characterized CIHP*) is based on the CIHP images [6]. We have kept the partition of the 33,280 images into 22,280 images for training and 5,000 for validation and testing. However, small changes have been made on human instance masks and the 19 semantic attribute classes. Moreover, 20 characteristic classes have been annotated in a pixel-wise way. The first and third rows of Fig. 1 show an example of the new CCIHP annotations. More examples can be found on the supplemental material. Fig. 2 shows the distribution of images per label.

Human instances 110,821 people (+121 humans compared to CIHP [6]) are annotated with full masks including accessories that did not have labels in the semantic attributes. Thus, belt or necklace are for example integrated into the human masks. The number of people per image is still around 3.



Fig. 2: Image distribution on the 19 (resp. 12, 4, 4) semantic attribute (resp. color, size, pattern) labels in CCIHP.

Semantic attributes We have modified the CIHP classes but we still get 19 classes of clothing and body parts in the end. 'Glasses' class now includes sunglasses, glasses and eyewears. 'Scarf' class is extended to all clothing worn around the neck: tie, bow-tie, scout scarf, ... Masks of the 'Dress' class are split into the 'UpperClothes' and 'Skirt' classes. Masks of 'Hair' class are augmented with facial hair like beard and mustache. Finally, a new 'Mask' class is added for all facial masks.

Size characteristics Four size classes characterize clothing attributes and/or hair: 'Short/small', 'Long/large', 'Undetermined' (when the attribute in truncated or occluded by another one), 'Sparse/bald' (for 'Hair' class only). Combining attribute classes with size characteristic gives fine-grained attribute labels. E.g., 'Pants' + 'Short' = shorts; 'Shoes' + 'Long' = boots.

Pattern characteristics Four pattern classes characterize clothing attributes and hair: 'Solid', 'Geometrical', 'Fancy', 'Letters'. The 'Solid' class is 3 times more represented than the others (cf. Figure 2) as all 'Hair' labels and lots of clothing are solid.

Color characteristics Twelve color classes characterize clothing attributes and hair: 'Brown', 'Red', 'Pink', 'Yellow', 'Orange', 'Green', 'Blue', 'Purple', 'Multicolor' (when several colors are evenly represented on the attribute part), and colors with no specific hue, 'Dark' (black to gray), 'Medium' (gray), 'Light' (gray to white). E.g.: 'Glasses' + 'Dark' = sunglasses; 'Glasses' + 'Medium' = reading glasses.

3. PROPOSED METHOD

With this new dataset, we propose an original baseline called HPTR for *Human Parsing with TRansformers*. Our approach is bottom-up and multi-task, sharing features between the different tasks to be scalable.

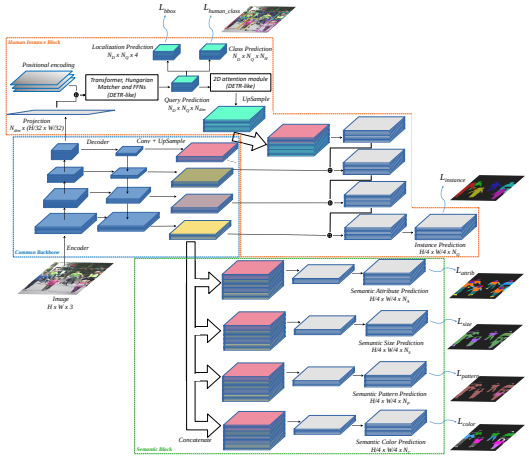


Fig. 3: HPTR overview. From an RGB HxW image, our multi-task bottom-up model predicts semantic maps with N_A (resp. N_S, N_P, N_C) channels for the N_A (resp. N_S, N_P, N_C) learned attribute (resp. size, pattern, color) classes. With the use of transformers, human instance maps are also predicted with N_Q channels. N_Q is also the number of queries in the transformers's decoder. N_{dim}, N_D are hidden transformer dimensions set as in DETR [24].

HPTR Architecture can be split into 3 blocks (cf. Fig. 3).
(1) The *Common Backbone* will share its features with all the task branches. It is composed of an encoder and a decoder to keep information at multiple resolutions, like in detector architectures. As in PandaNet 3D human pose estimation model [25], each pyramid feature map of the decoder is passed through 4 convolutional layers and resized to the size of the highest resolution level. These feature maps feed the different task branches of the second and third blocks.
(2) In the *Human Instance Block*, we exploit transformers for human detection as in the recent DETR approach (see details in [24]), to ensure that our architecture keeps bottom-up property and does not require post-processing at the end of the inference. However, the mask branch of our *Human Instance Block* differs from DETR: To share a maximum of features with the other branches of our model, we do not generate instance mask from feature maps of the encoder. Instead, we concatenate the up-scaled feature maps of the decoder with the feature maps from the transformer's outputs. It is achieved by creating a top-down pathway with lateral connections to bottom-up convolutional layers. In the end, this block provides human bounding boxes, scores and masks.
(3) *Semantic Block* corresponds to the semantic attribute and characteristic (size, pattern, color) branches. This block allows a pixel-wise segmentation of people body parts, clothing and their characteristics. *Semantic Block* is made of 4 similar branches, one per task. Each of them is fed with the concatenation of the up-scaled pyramid feature maps of the decoder. These maps are forwarded into a convolutional semantic head to predict the semantic attribute, size, pattern and color outputs according to the classes defined in CCIHP.

Training objective To train the *Human Instance Block*, we follow DETR [24] and use the same objective. The training loss $\mathcal{L}_{human}$ is the sum of $\mathcal{L}_{Hungarian}$ (not presented in Fig. 3), $\mathcal{L}_{bbox}$ (an l_1 loss), $\mathcal{L}_{human_class}$ (a cross-entropy loss), and $\mathcal{L}_{instance}$ (composed of a DICE loss [26] and a Focal loss [27]). In the *Semantic Block*, we also use for each head a DICE loss and a Focal loss (rather than a cross-entropy loss) to better deal with imbalanced classes. No weighting is used between each loss. The global training objective is then as follows: $\mathcal{L} = \mathcal{L}_{human} + \mathcal{L}_{attrib} + \mathcal{L}_{size} + \mathcal{L}_{pattern} + \mathcal{L}_{color}$.

4. EXPERIMENTS AND RESULTS

Implementation Details HPTR has been implemented in Pytorch. The encoder is a ResNet50 pre-trained on ImageNet dataset. The *Human Instance* and *Semantic Blocks* are trained jointly for 300 epochs on 8 Titan X (Pascal) GPUs with 1 image per batch. The long side of training and validation images can not exceed 512 pixels. N_Q is set to 40 person queries per image. Please refer to [24] for other hyper-parameter settings and input transformations.

Datasets and Evaluation Protocols HPTR is evaluated on proposed dataset CCIHP, as a first baseline of characterized multi-HP, and also on CIHP [6] for comparison with the state of the art of multi-HP without characterization. We follow PGN [6] evaluation implementation by using *mean Intersection over Union (mIoU)* [28] for attribute and characteristic (size, pattern, color) semantic segmentation evaluation. As instance-level HP evaluation metric, we use *mean Average Precision based on region* (AP^r_{vol}) [29]. We also compute *mean Average Precision based on part* (AP^p_{vol}) [17] following Parsing R-CNN [23] evaluation protocol. Finally, we extend AP^r_{vol} metric to evaluate attribute characterization: Whereas

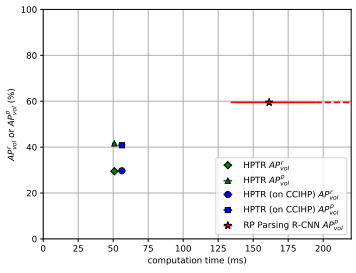


Fig. 4: Precision/time trade-off of HPTR on CIHP and CCIHP (on a TITAN X GPU), compared to RP Parsing R-CNN that achieved best precision and speed in multi-HP state of the art. HPTR is the fastest method and has constant time while RP Parsing R-CNN (a top-down approach) is more accurate but slower and not scalable: red horizontal solid line shows time variation with 2 to 18 people per image.

	all	Hat	Hair	Glove	Sunglasses	Glasses	UpperClothes	Dress	Mask	Coat	Socks	Pants	Torso-skin	Scarf	Scarf/Tie	Skirt	Face	L-arm	R-arm	L-leg	R-Leg	L-shoe	R-shoe
CIHP mIoU	53.0	58.1	70.4	10.5	33.9	-	53.5	34.6	-	41.2	17.3	49.2	57.2	7.6	-	19.5	69.0	38.9	31.7	24.1	18.8	20.1	16.9
CCIHP mIoU	52.1	55.0	69.7	7.5	-	41.0	48.9	-	8.2	44.7	16.8	50.0	57.8	-	34.1	38.0	66.0	38.8	30.2	23.7	19.5	19.5	17.1
CCIHP AP^r_{vol}	29.7	52.0	60.9	6.6	-	32.1	45.1	-	10.7	48.3	12.8	52.8	48.1	-	28.0	40.3	71.3	9.0	11.6	10.8	8.4	7.2	7.2

(a)

	semantic color													semantic size					semantic pattern				
	all	Dark	Medium	Light	Brown	Red	Pink	Yellow	Orange	Green	Blue	Purple	Multicolor	all	Short	Long	Undet.	Sparse	all	Solid	Geom.	Fancy	Letters
mIoU	43.8	60.3	23.1	31.5	9.0	19.7	11.4	11.2	1.1	12.4	15.7	4.9	14.2	58.8	55.0	58.6	20.5	14.6	67.4	72.3	30.0	21.0	16.3
AP^{cr}_{vol}	15.0	40.8	18.4	25.2	5.5	17.4	9.1	9.6	0.07	14.8	22.3	3.2	13.2	24.5	33.1	37.5	13.5	13.7	20.9	36.9	14.4	14.1	18.2

(b)

Table 1: HPTR performances per class in %. (a) Results on CIHP and CCIHP attributes. (b) Results on CCIHP color, size and pattern characteristics (Undet. stands for undetermined and Geom. for Geometrical).

AP^r_{vol} evaluates the prediction of attribute (class & score) relative to each instanced attribute mask, *mean Average Precision based on characterized region (AP^{cr}_{vol})* evaluates the prediction of characteristic (class & score) relative to each instanced *and* characterized attribute mask, *independently* of the attribute class prediction. Thus, AP^{cr}_{vol} is jointly conditional on human instance segmentation, attribute delineation and characteristic segmentation.

Comparison with state-of-the-art of multi-HP without characterization As multi-HP methods do not manage attribute characterization proposed in CCIHP, we first evaluate HPTR without characterization task, on CIHP dataset [6], to give an idea of its speed vs precision trade-off relative to these methods. We compare HPTR with best methods of each approach family that have available models. Inference time is averaged after 50 runs on a Titan X GPU, using CIHP images containing from 2 to 18 people. *Top-down* approaches reach the state-of-the art precision thanks to their local parsing branches: *two-stage* M-CE2P [4] reports AP^r_{vol} = 42.8% and *one-stage* RP Parsing R-CNN [5] gets AP^p_{vol} = 59.5%. However, top-down approaches are known to have computation times dependent on the number of people in a scene and not be easily scalable. Typically, on CIHP images, M-CE2P (resp. RP Parsing R-CNN) runs in 752 ms–6.6 s (resp. 136–195 ms) according to the number (2–18) of people per image. By extrapolation to 40 people, this time would go beyond 14 s for M-CE2P and 285 ms for RP Parsing R-CNN. In contrast, *bottom-up* approaches, generally less accurate, have the advantage to run in constant time, independent of the number of people. NAN [17] runs in about 275 ms (but no AP on CIHP was reported). As for PGN [6], it reaches AP^r_{vol} = 33.6%, and AP^p_{vol} = 39.0%, in around 1.4 s without counting additional post-processing time. These models are more than 5 and 29 times slower than HPTR. Indeed, our bottom-up approach has a good speed/precision trade-off with a low constant time of around 50 ms and precision similar to PGN (AP^r_{vol} = 29.5%, AP^p_{vol} = 41.6%). Fig. 4 shows the speed/precision trade-off for the most precise method (RP Parsing RCNN) and the proposed HPTR which is the fastest method of the state of the art while having precision comparable to the most precise bottom-up method (PGN). Low and constant computation time is essential when applying multi-HP on large amounts of possibly crowded videos.

Results of multi-HP with characterization Now, HPTR is evaluated on CCIHP and gets an overall mIoU of 52.1%, AP^r_{vol} of 29.7% and AP^p_{vol} of 40.8% (cf. Tab. 1a for results detailed per attribute class). mIoU results are also presented for CIHP with small differences in classes. Per class results are close to those obtained on CIHP. The main differences are on 'Glasses', 'Scarf/Tie' and 'Skirt' classes that are more represented in CCIHP (cf. Sec. 2). Tab. 1b shows the mIoU and our new metric AP^{cr}_{vol} for all characteristic classes. Overall, HPTR reaches an mIoU of 43.8% (resp. 58.8% and 67.4%) and an AP^{cr}_{vol} of 15.0% (resp. 24.5% and 20.9%) for the color (resp. size and pattern) labels. We can see that results are directly correlated to class frequency (cf. Fig. 2). So we would like to address this class imbalance issue as future work to improve HPTR performance. Consistency between characteristic and attribute masks, observed qualitatively as in Fig. 1, is driven by the backbone features shared by the task branches. Besides, computation time of 56 ms shows that HPTR is also scalable with these 3 additional characterization tasks (compared to 50 ms without characterization, cf. Fig. 4). Thus, the use of global branches for additional tasks, instead of local branches, gives another advantage over top-down approaches.

5. CONCLUSION

In this article, we propose CCIHP, the first multi-HP dataset with systematic characterization of instance-level attributes. It is based on CIHP images, the largest existing multi-HP in-the-wild dataset. We have defined 20 classes of characteristics split into 3 categories (size, pattern and color) to better describe each human attribute. To learn these characteristics, we have developed HPTR, a bottom-up, and multi-task baseline. It has low and constant computation time. Thus, it is scalable with the number of people per image and the number of tasks to learn. We hope that research towards fast and accurate methods for more complete human descriptions will be encouraged thanks to this new dataset and baseline.

Acknowledgments This publication was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council.

6. REFERENCES

[1] A. Loesch, J. Rabarisoa, and R. Audigier, "End-to-end person search sequentially trained on aggregated dataset," in *ICIP*, 2019.

[2] Y. Lin, L. Zheng, Z. Wu, Y. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *PR*, vol. 95, pp. 151–161, 2019.

[3] K. Gong, Y. Gao, X. Liang, X. Shen, M. Yang, and L. Lin, "Graphonomy: Universal human parsing via graph transfer learning," in *CVPR*, 2019.

[4] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, "Devil in the details: Towards accurate single and multiple human parsing," in *AAAI*, 2019.

[5] L. Yang, Q. Song, Z. Wang, M. Hu, C. Liu, X. Xin, and S. Xu, "Renovating parsing r-cnn for accurate multiple human parsing," in *ECCV*, 2020.

[6] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *ECCV*, 2018.

[7] C. W. Hsieh, C. Y. Chen, C. L. Chou, H. H. Shuai, J. Liu, and W. H. Cheng, "FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information," in *ACM ICM*, 2020.

[8] W. H. Cheng, S. Song, C. Y. Chen, S. C. Hidayati, and J. Liu, "Fashion meets computer vision: A survey," in *arXiv*, 2020.

[9] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan, "A deformable mixture parsing model with parselets," in *ICCV*, 2013.

[10] K. Yamaguchi, M. Hadi Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *ICCV*, 2013.

[11] W. Yang, P. Luo, and L. Lin, "Clothing co-parsing by joint image segmentation and labeling," in *CVPR*, 2014.

[12] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, "Fashion parsing with weak color-category labels," *TM*, vol. 16, no. 1, pp. 253–265, 2013.

[13] S. Zheng, F. Yang, M. H.. Kiapour, and R. Piramuthu, "Modanet: A large-scale street fashion dataset with polygon annotations," in *ACM ICM*, 2018.

[14] 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.

[15] F. Xia, P. Wang, X. Chen, and A. L. Yuille, "Joint multi-person pose estimation and semantic part segmentation," in *CVPR*, 2017.

[16] J. Li, J. Zhai, Y. Wei, C. Lang, Y. Li, T. Sim, S. Yan, and J. Feng, "Multiple-human parsing in the wild," in *arXiv*, 2017.

[17] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *ACM ICM*, 2018.

[18] H. He, J. Zhang, Q. Zhang, D. Tao, M. Yang, and L. Lin, "Grapy-ML: graph pyramid mutual learning for cross-dataset human parsin," in *AAAI*, 2020.

[19] Q. Li, A. Arnab, and P. H. Torr, "Holistic, instance-level human parsing," in *arXiv*, 2017.

[20] R. Ji, D. Du, L. Zhang, L. Wen, Y. Wu, C. Zhao, F. Huang, and S. Lyu, "Learning semantic neural tree for human parsing," in *arXiv*, 2019.

[21] X. Liu, M. Zhang, W. Liu, J. Song, and T. Mei, "Braidnet: Braiding semantics and details for accurate human parsing," in *ACM ICM*, 2019.

[22] H. Qin, W. Hong, W. C. Hung, Y. H. Tsai, and M. H. Yang, "A top-down unified framework for instance-level human parsing," in *BMVC*, 2019.

[23] L. Yang, Q. Song, Z. Wang, and M. Jiang, "Parsing r-cnn for instance-level human analysis," in *CVPR*, 2019.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.

[25] A. Benzine, F. Chabot, B. Luvison, Q.C. Pham, and C. Achard, "Pandanet: Anchor-based single-shot multi-person 3d pose estimation," in *CVPR*, 2020.

[26] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*, 2016.

[27] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, T. Sim, S. Yan, and J. Feng, "Focal loss for dense object detection," in *ICCV*, 2017.

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[29] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *ECCV*, 2014.