DE LA RECHERCHE À L'INDUSTRIE

DESCRIBE ME IF YOU CAN!

CHARACTERIZED INSTANCED-LEVEL HUMAN PARSING

Angélique LOESCH – Romaric AUDIGIER

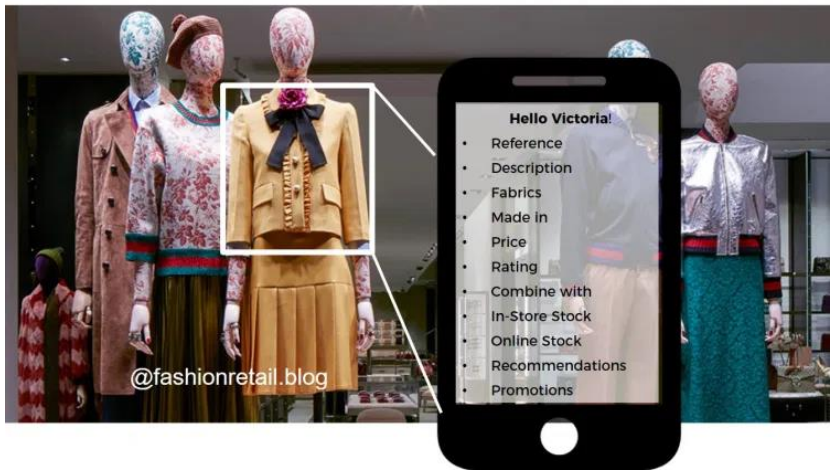
{angelique.loesch, romaric.audigier}@cea.fr

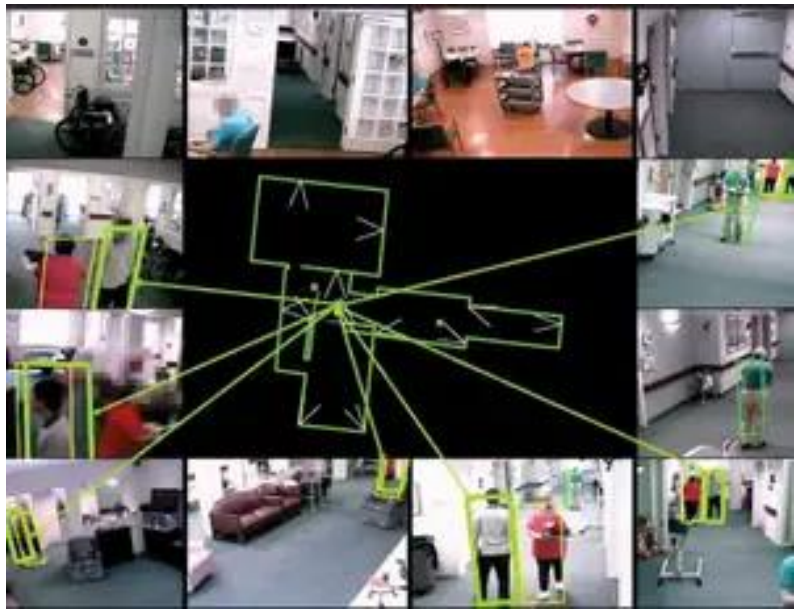
CEA LIST, Vision and Learning Lab for Scene Analysis, Université Paris-Saclay, France / Vision Lab, Thales SIX GTS, France

Human Parsing

Human semantic description with extraction of semantic attributes

- **Useful for**
  - image content description,
  - image generation for virtual reality applications,
  - person retrieval from a natural-description query, for security applications…





Source: Shoou-I Yu, Yi Yang and Alexander Hauptmann/Carnegie Mellon University

Human Parsing

Human semantic description with extraction of semantic attributes

- **Useful for**
  - image content description,
  - image generation for virtual reality applications,
  - person retrieval from a natural-description query, for security applications…

- **Many datasets for fashion applications but single-person images in controlled environments**

Single-person dataset examples: ATR [Liang15], Modanet [Zheng18], Deepfashion2 [Ge19]

Human Parsing

Human semantic description with extraction of semantic attributes

- **Useful for**
  - image content description,
  - image generation for virtual reality applications,
  - person retrieval from a natural-description query, for security applications…
- **Many datasets for fashion applications but single-person images in controlled environments**
- **Datasets in-the-wild multi-person datasets but without attribute qualification**

Multi-person human parsing dataset examples: MHP [Li17], CIHP [Gong18]

Characterized Human Parsing: Contributions

Original image

Person Segmentation

Human instances

Human Parsing

Characterized Human Parsing

1 2

Attributes

Size

Color

Pattern

Source: RGB Images from CIHP / COCO datasets

1 CCIHP

Characterized **C**rowd **I**nstance-level **H**uman **P**arsing

A new Human Parsing dataset with characteristics

- **Multi-person**
- **In the wild** scenes
- Addition of a **new dimension of attribute analysis**
  - Qualification of attributes through finely annotated characteristics

2 HPTR

Human **P**arsing with **TR**ansformers

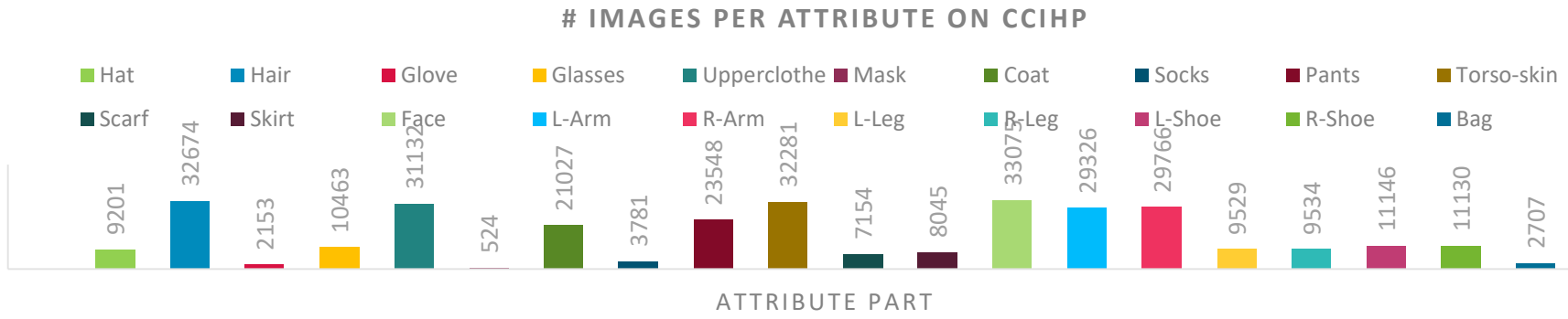
A new model based on transformers

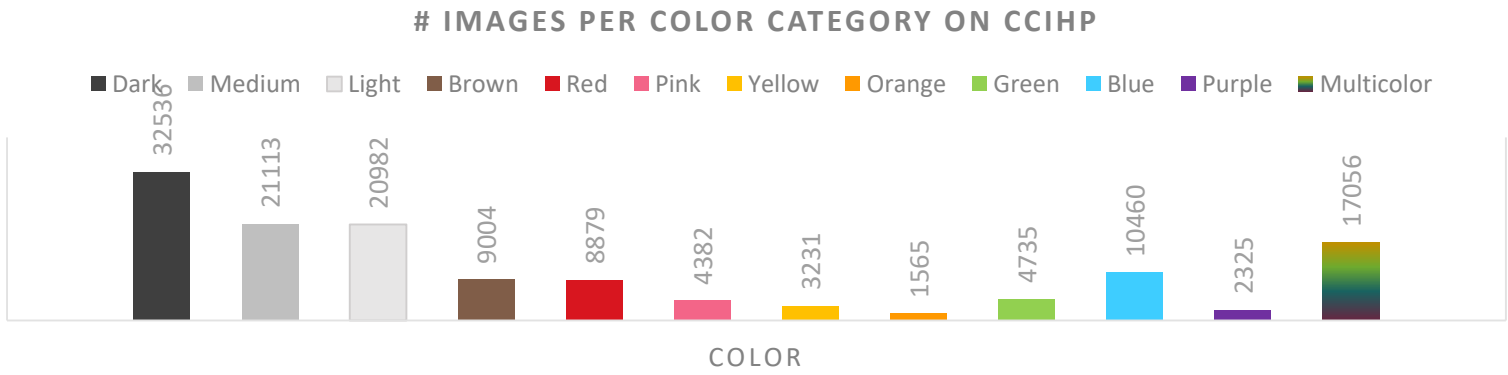
- **bottom-up**
- **multi-task**
- No post-processing needed
- Low constant processing time
  - Scalability for a system deployment

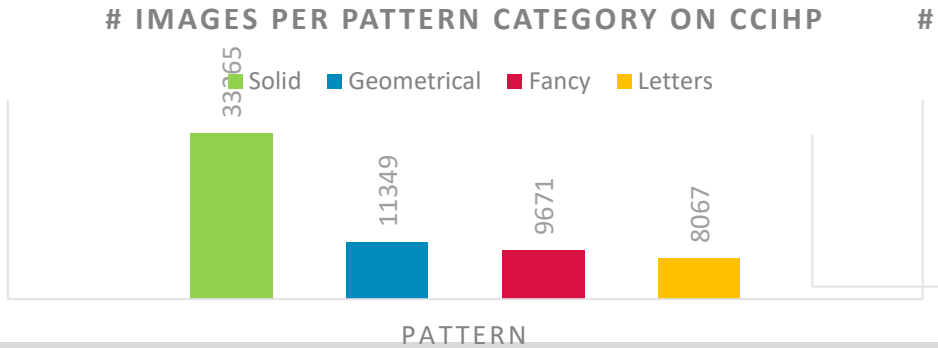
Proposed dataset: CCIHP

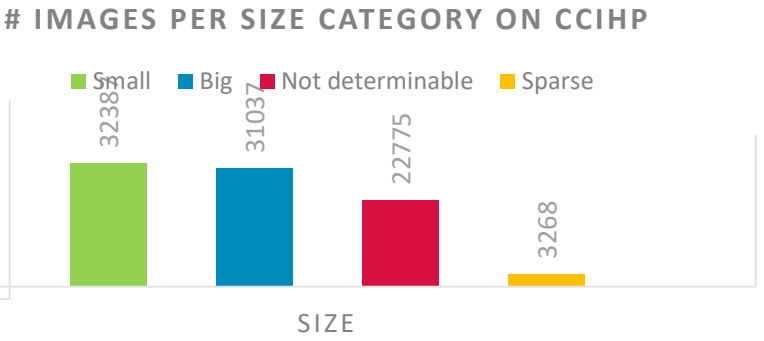
- **Characterized Crowd Instance-level Human Parsing dataset**
  - 33,280 images
  - 110,821 persons
  - Based on CIHP dataset [Gong18]

- **First dataset of its kind**
  - In the wild scenes
  - Multiple persons per image
  - Pixel-wise annotation
  - New taxonomy
    - 20 attribute classes
    - 20 characterization classes
      - Color
      - Pattern
      - Size







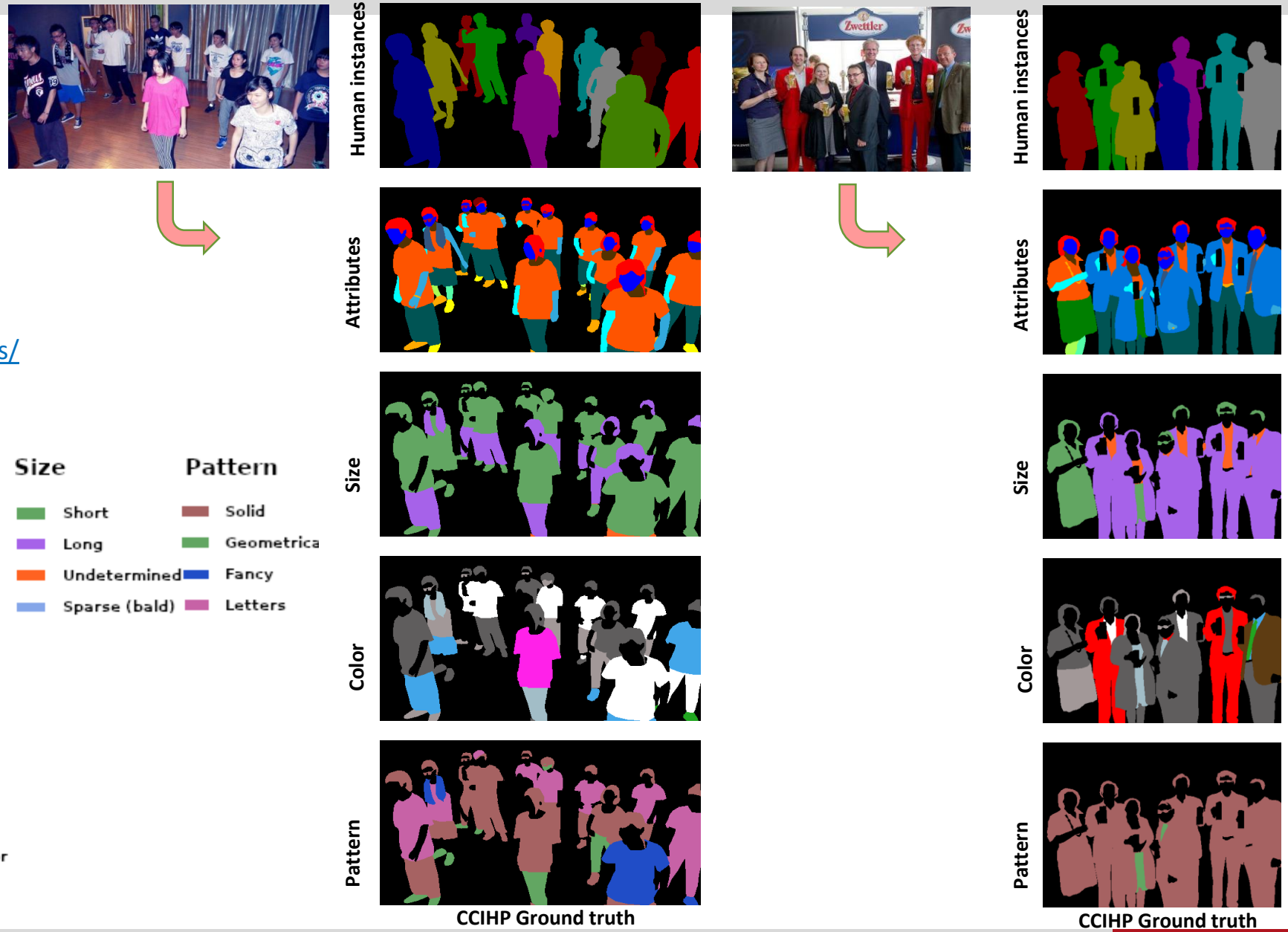


Pixano: https://pixano.cea.fr/

Proposed dataset: CCIHP

Available on
https://kalisteo.cea.fr/index.php/free-resources/

Attribute				Color	Size	Pattern
Hat	Scarf/Tie			Dark	Short	Solid
Hair	Skirt			Medium	Long	Geometrica
Glove	Face			Light	Undetermined	Fancy
(Sun)Glasses	L-arm			Brown	Sparse (bald)	Letters
UpperClothes	R-arm			Red		
Mask	L-leg			Pink		
Coat	R-leg			Yellow		
Socks	L-shoe			Orange		
Pants	R-shoe			Green		
Torso-skin				Blue		
				Purple		
				Multicolor		



Source: Images from CIHP / COCO datasets

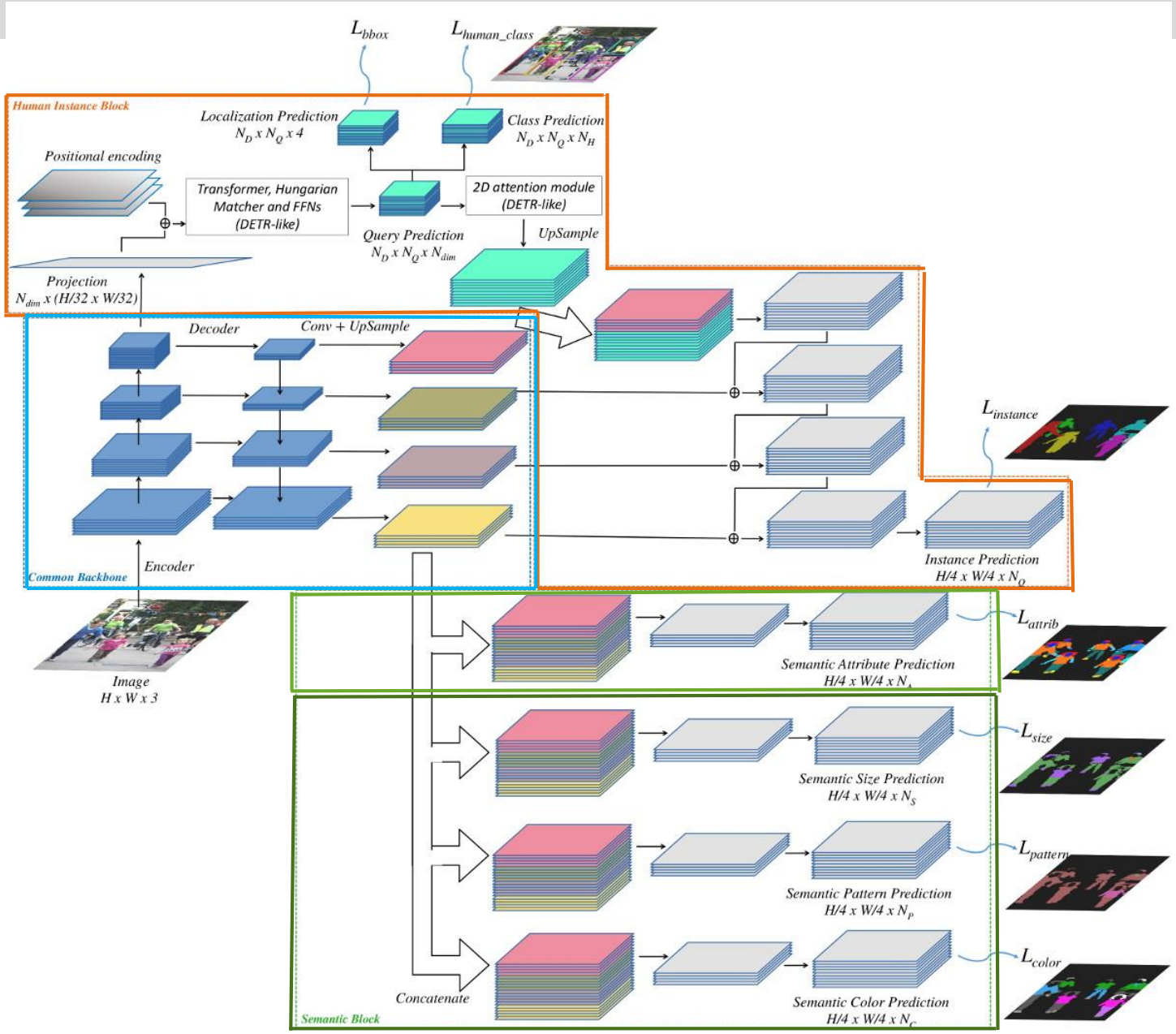
Proposed method: HPTR

- **Human Parsing with TRansformers**
- **Bottom-up multi-task model with a shared backbone and 3 main blocks trained simultaneously**

  - **A human detection and instance segmentation** based on DETR [Carion20]
  → prediction of bounding boxes, confidence score and instance segmentation mask

  - **An attribute semantic segmentation**
  → prediction of the semantic masks of the attributes

  - **Characteristic semantic segmentations**
  → prediction of the semantic masks of the 3 types of characterization (color, size, pattern)



$$L = L_{human} + L_{attrib} + L_{size} + L_{pattern} + L_{color}$$

Proposed method: HPTR

Best trade-off computation time / accuracy

- **On CIHP dataset.**
  - HPTR is the fastest (3x) method and has constant time
  - HPTR is competitive with SOTA bottom-up methods
  - RP Parsing R-CNN (top-down approach) is more accurate but slower and not scalable
- **On CCIHP dataset (ours)**
  - New metric AP^{cr}_{vol}: mean Average Precision based on characterized region
    - prediction of characteristic (class & score) relative to each instanced and characterized attribute mask, independently of the attribute class prediction
  - HPTR is also constant and low with these 3 additional characterization tasks

Family	Methods	CCIHP validation set (ours)						
		AP^{p}_{vol} (%)	AP^{r}_{vol} (%)	Size AP^{cr}_{vol} (%)	Pattern AP^{cr}_{vol} (%)	Color AP^{cr}_{vol} (%)	Inf. time (ms)	
							2 people	18 people
Bottom-up	HPTR (ours)	40.8	29.7	24.5	20.9	15.0	56	56

Family	Methods	CIHP validation set [Gong18]			
		AP^{p}_{vol} (%)	AP^{r}_{vol} (%)	Inf. time (ms)	
				2 people	18 people
Top-down	M-CE2P [Ruan19]	-	42.8	752	6600
	Parsing R-CNN [Yang20]	59.5	-	136	195
Bottom-up	PGN [Gong18]	39.0	**33.6**	1400	1400
	NAN [Zhao18]	-	-	275	275
	HPTR (ours)	**41.6**	29.5	**50**	**50**

Precision/time trade-off on CCIHP (left) and CIHP (right) (on a Titan X GPU): SOTA comparison

Proposed method: HPTR

Qualitative results

Source: Images from CIHP / COCO datasets

Conclusion

- **CCIHP**, the first multi-HP dataset with systematic characterization of instance-level attributes
- **HPTR**, a bottom-up, and multi-task baseline, with low constant processing time, whatever the number of people per image
- We hope that research towards fast and accurate methods for more complete human descriptions will be encouraged thanks to this new dataset and baseline
- Please check our CCIHP dataset on **https://kalisteo.cea.fr/index.php/free-resources/**

References

- **[Carion20]** Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European Conference on Computer Vision* (pp. 213-229). Springer, Cham.

- **[Ge19]** Ge, Y., Zhang, R., Wang, X., Tang, X., & Luo, P. (2019). Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5337-5345).

- **[Gong18]** Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., & Lin, L. (2018). Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 770-785).

- **[Li17]** Li, J., Zhao, J., Wei, Y., Lang, C., Li, Y., Sim, T., ... & Feng, J. (2017). Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*.

- **[Liang15]** Liang, X., Liu, S., Shen, X., Yang, J., Liu, L., Dong, J., ... & Yan, S. (2015). Deep human parsing with active template regression. *IEEE transactions on pattern analysis and machine intelligence*, *37*(12), 2402-2414.

- **[Yang20]** Yang, L., Song, Q., Wang, Z., Hu, M., Liu, C., Xin, X., ... & Xu, S. (2020, August). Renovating parsing R-CNN for accurate multiple human parsing. In *European Conference on Computer Vision* (pp. 421-437). Springer, Cham.

- **[Zhao18]** Zhao, J., Li, J., Cheng, Y., Sim, T., Yan, S., & Feng, J. (2018, October). Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *Proceedings of the 26th ACM international conference on Multimedia* (pp. 792-800).

- **[Zheng18]** Zheng, S., Yang, F., Kiapour, M. H., & Piramuthu, R. (2018, October). Modanet: A large-scale street fashion dataset with polygon annotations. In *Proceedings of the 26th ACM international conference on Multimedia* (pp. 1670-1678).

Thank you for watching

Acknowledgments:

- **ANR TeamSports project**
- **FactoryIA supercomputer supported by the Ile-de-France Regional Council**