Transferring Dense Pose to Proximal Animal Classes

Artsiom Sanakoyeu*
Heidelberg University

Vasil Khalidov
Facebook AI Research

Maureen McCarthy
MPI for Evolutionary Anthropology

Andrea Vedaldi
Facebook AI Research

Natalia Neverova
Facebook AI Research

In Section 1 we provide more details on our implementation of the Multi-head R-CNN network. Then, in Section 3 we describe additional ablation studies on the advantages of the auto-calibrated training, as well as other architectural choices. Finally, Section 4 refers the reader to the qualitative results obtained on videos from the Chimp&See dataset.

1. Architecture

We introduced a number of changes and improvements in the DensePose head of the standard DensePose R-CNN architecture of [2] with ResNet-50 [4] backbone. These changes are listed below for the affected branches; other branches remained unchanged and correspond exactly to the Mask R-CNN architecture of [3].

- We have increased the RoI resolution from 14×14 to 28×28 in the DensePose head, as proposed in [7].
- We have replaced the 8-layer DensePose head with the geometric and context encoding (GCE) module [7], combining a non-local convolutional layer [6] with the atrous spatial pyramid pooling (ASPP) [1].
- We have replaced the original FPN of DensePose R-CNN with a Panoptic FPN [5].

Each of these modifications led to increase in network performance due to improved multi-scale context aggregation. We refer the reader to the work of [7] for ablation studies whose results are aligned well with our own observations.

To predict or we simply extend the output layer of the corresponding head by doubling the number of its neurons.

Our codebase, network configuration files for each experiment and pretrained models will be publicly released.

2. Computational cost

Our auto-calibrated model has a negligible computational overhead ($< 1\%$) compared to the baseline model. Before training the *student*, sampling of the pseudo-labels requires one forward pass of the *teacher* network over the unlabeled dataset. The *teacher* and the *student* networks share the same architecture.

3. Ablation studies

First, we report performance of the original Mask R-CNN [3] framework, as well as our auto-calibrated version of the same architecture, on detection and segmentation tasks (see Tab. 1). Training in the auto-calibration setting resulted in minor gains on the COCO dataset that the model was trained on, but, as expected, led to major improvements in performance on the out-of-distribution data (DensePose-Chimps and Chimp&See).

Second, Tab. 2 shows results of replacing the proposed binary foreground-background segmentation in the DensePose head (a) with 15-way coarse body part segmentation as in the original DensePose-RCNN framework [2] (b). We can see that binary segmentation generalizes better than the 15-way. We have also experimented with using the binary mask from the Mask R-CNN head instead of mask produced by the DensePose head (Tab. 2 (c)) *during inference step*. Moreover, even though exploiting the mask from the separate mask head at test time results in better performance, complete removal of the mask from the DensePose head leads to under-training and decreased accuracy of estimation of uv-coordinates (since in this case the DensePose head receives only sparse supervisory signals at the annotated locations).

4. Qualitative results

In addition, we also point the readers to the video samples* from the Chimp&See dataset showing frame-by-frame predictions produced by our model before (*teacher*) and after self-training (*student*). The results produced by the *student* network are generally significantly more stable.

* https://asanakoy.github.io/densepose-evolution

	COCO `minival`		DensePose-Chimps		Chimp&See	
model	**AP**$_D$	**AP**$_S$	**AP**$_D$	**AP**$_S$	**AP**$_D$	**AP**$_S$
Mask RCNN	40.98	**37.17**	48.3	44.92	40.56	33.91
σ-Mask RCNN	**41.12** (+0.14)	37.09 (-0.08)	**52.05** (+3.75)	**47.94** (+3.02)	**42.9** (+2.34)	**34.74** (+0.82)

Table 1: **Auto-calibrated Mask R-CNN [3]:** detection, instance segmentation on COCO `minival` (all classes).

	model	Mask in DensePose head	**AP**	**AP**$_{50}$	**AP**$_{75}$
a)	DensePose-RCNN* (σ)	binary	53.20	88.27	56.98
b)	DensePose-RCNN* (σ)	15-way	50.87	86.91	54.49
c)	DensePose-RCNN* (σ) + mask from the mask head	binary	**54.35**	**88.58**	**60.28**

Table 2: **Ablation study of the mask in the DensePose head.** Reports the DensePose performance on DensePose-COCO `minival`. a) our proposed architecture; b) replace the binary segmentation of the DensePose head with 15-way coarse body part segmentation as in the original DensePose-RCNN framework [2]; c) use the binary mask from the DensePose head during training, but substitute it with the mask from the separate mask head during inference.

5. Acknowledgements

We thank all parties performing or supporting collection of the Chimp&See dataset, including:

(a) individual contributors: Theophile Desarmeaux, Kathryn J. Jeffery, Emily Neil, Emmanuel Ayuk Ayimisin, Vincent Lapeyre, Anthony Agbor, Gregory Brazzola, Floris Aubert, Sebastien Regnaut, Laura Kehoe, Lucy DAuvergne, Nuria Maldonado, Anthony Agbor, Emmanuelle Normand, Virginie Vergnes, Juan Lapuente, Amelia Meier, Juan Lapuente, Alexander Tickle, Heather Cohen, Jodie Preece, Amelia Meier, Juan Lapuente, Roman M. Wittig, Dervla Dowd, Sorrel Jones, Sergio Marrocoli, Vera Leinert, Charlotte Coupland, Villard Ebot Egbe, Anthony Agbor, Volker Sommer, Emma Bailey, Andrew Dunn, Inaoyom Imong, Emmanuel Dilambaka, Mattia Bessone, Amelia Meier, Crickette Sanz, David Morgan, Aaron Rundus, Rebecca Chancellor, Felix Mulindahabi, Protais Niyigaba, Chloe Cipoletta, Michael Kaiser, Kyle Yurkiw, Bradley Larson, Alhaji Malikie Siaka, Liliana Pacheco, Manuel Llana, Henk Eshuis, Erin G. Wessling, Mohamed Kambi, Parag Kadam, Alex Piel, Fiona Stewart, Katherine Corogenes, Klaus Zuberbuehler, Kevin Lee, Samuel Angedakin, Kevin E. Langergraber, Christophe Boesch, Hjalmar Kuehl, Mimi Arandjelovic, Paula Dieguez, Mizuki Murai, Yasmin Moebius, Joana Pereira, Silke Atmaca, Kristin Havercamp, Nuria Maldonado, Colleen Stephens;

(b) funding agencies: Max Planck Society, Max Planck Society Innovation Fund, Heinz L. Krekeler Foundation;

(c) ministries and governmental organizations: Agence Nationale des Parcs Nationaux (Gabon), Centre National de la Recherche Scientifique (CENAREST) (Gabon), Conservation Society of Mbe Mountains (CAMM) (Nigeria), Department of Wildlife and Range Management (Ghana), Direction des Eaux, Forłts et Chasses (Senegal), Eaux et Forets (Mali), Forestry Commission (Ghana), Forestry Development Authority (Liberia), Institut Congolais pour la Conservation de la Nature (DR-Congo), Instituto da Biodiversidade e das reas Protegidas (IBAP), Makerere University Biological Field Station (MUBFS) (Uganda), Ministere de lEconomie Forestiere (R-Congo), Ministere de la Recherche Scientifique et de l'Innovation (Cameroon), Ministere de la Recherche Scientifique (DR-Congo), Ministere de l'Agriculture de l'Elevage et des Eaux et Forets (Guinea), Ministere de le Recherche Scientifique et Technologique (R-Congo), Ministere des Eaux et Forets (Cote dIvoire), Ministere des Forets et de la Faune (Cameroon), Ministre de l'Environnement et de l'Assainissement et du Developpement Durable du Mali, Ministro da Agricultura e Desenvolvimento Rural (Guinea-Bissau), Ministry of Agriculture, Forestry and Food Security (Sierra Leone), Ministry of Education (Rwanda), National Forestry Authority (Uganda), National Park Service (Nigeria), National Protected area Authority (Sierra Leone), Rwanda Development Board (Rwanda), Socit Equatoriale dExploitation Forestire (SEEF) (Gabon), Tanzania Commission for Science and Technology (Tanzania), Tanzania Wildlife Research Institute (Tanzania), Uganda National Council for Science and Technology (UNCST), (Uganda), Uganda Wildlife Authority (Uganda);

(d) non-governmental organizations: Budongo Conservation Field Station (Uganda), Ebo Forest Research Station (Cameroon), Fongoli Savanna Chimpanzee Project (Senegal), Foundation Chimbo (Boe), Gashaka Primate Project (Nigeria), Gishwati Chimpanzee Project (Rwanda), Goualougo Triangle Ape Project, Jane Goodall Insitute Spain (Dindefelo) (Senegal), Korup Rainforest

Conservation Society (Cameroon), Kwame Nkrumah University of Science and Technology (KNUST) (Ghana), Loango Ape Project (Gabon), Lukuru Wildlife Research Foundation (DRC), Ngogo Chimpanzee Project (Uganda), Nyungwe-Kibira Landscape, Rwanda-Burundi (WCS), Projet Grands Singes, La Belgique, Cameroon (KMDA), Station dEtudes des Gorilles et Chimpanzees (Gabon), Tai Chimpanzee Project (Cote dIvoire), The Aspinall Foundation, (Gabon), Ugalla Primate Project (Tanzania), WCS (Conkouati-Douli NP) (R-Congo), WCS Albertine Rift Programme (DRC), Wild Chimpanzee Foundation (Cote d'Ivoire), Wild Chimpanzee Foundation (Guinea), Wild Chimpanzee Foundation (Liberia), Wildlife Conservation Society (WCS) Nigeria (Nigeria), WWF (Campo Maan NP) (Cameroon), WWF Congo Basin (DRC).

References

[1] L. Chen, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convlution for semantic image segmentation. *ECCV*, 2018. 1

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. *CVPR*, 2018. 1, 2

[3] K. He, G. Gkioxari, and P. Dollár and. R. Girshick. Mask R-CNN. *ICCV*, 2017. 1, 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. *CVPR*, pages 6399–6408, 2019. 1

[6] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. *CVPR*, 2018. 1

[7] Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing r-cnn for instance-level human analysis. *CVPR*, 2018. 1