Universitatea Transilvania din Brașov
FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

Universitatea Transilvania din Brașov

29.01.2019

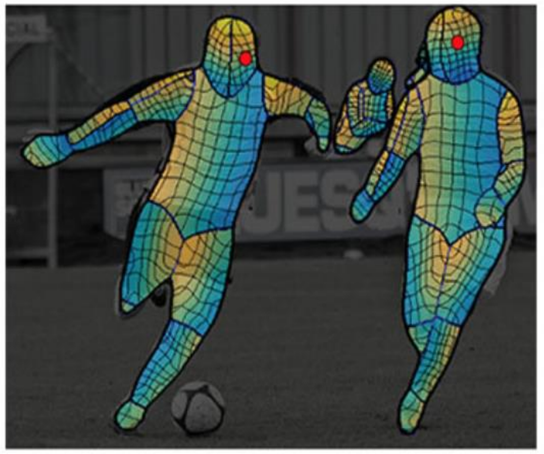
PRESENTED BY:
Ionescu Vlad

Dense Pose

Facebook research

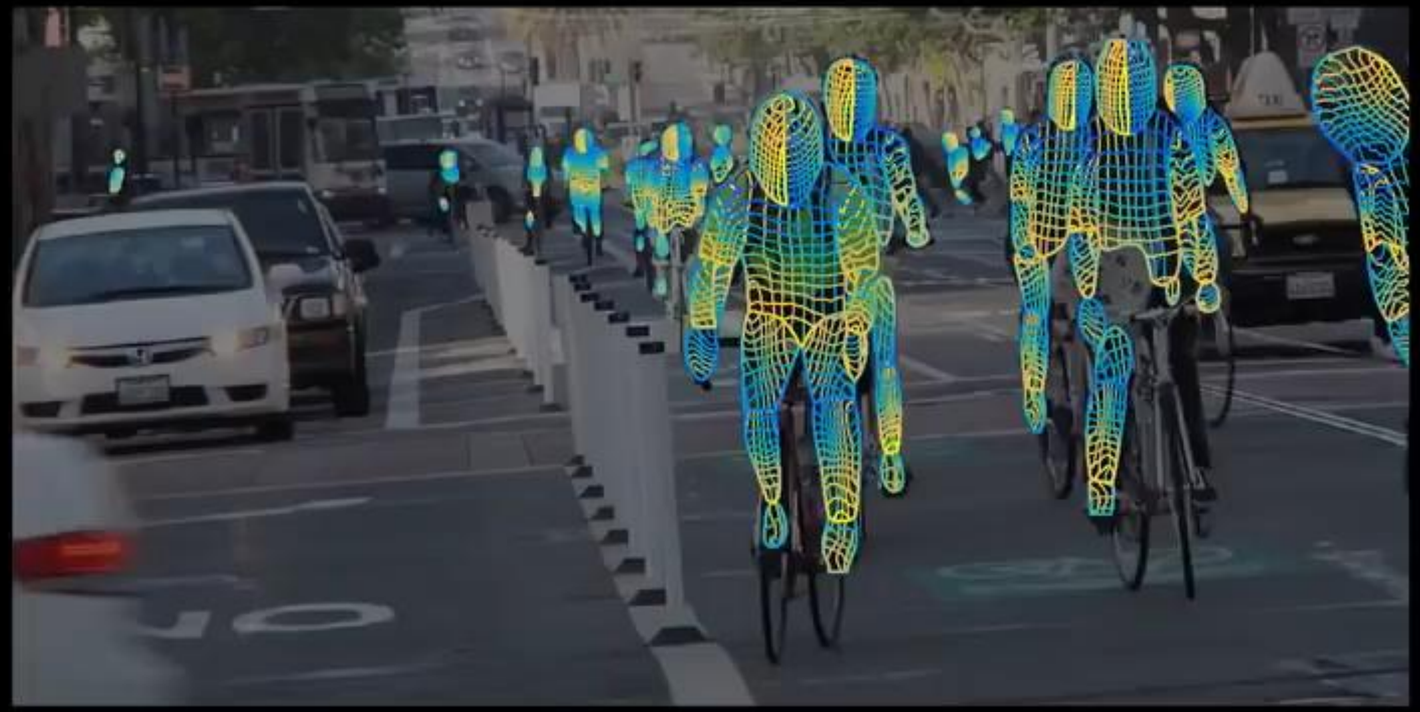
What is Dense Pose

Dense human pose estimation aims at mapping all human pixels of an RGB image to the 3D surface of the human body.



DensePose:

Dense Human Pose Estimation In The Wild



Rıza Alp Güler *
INRIA, CentraleSupélec

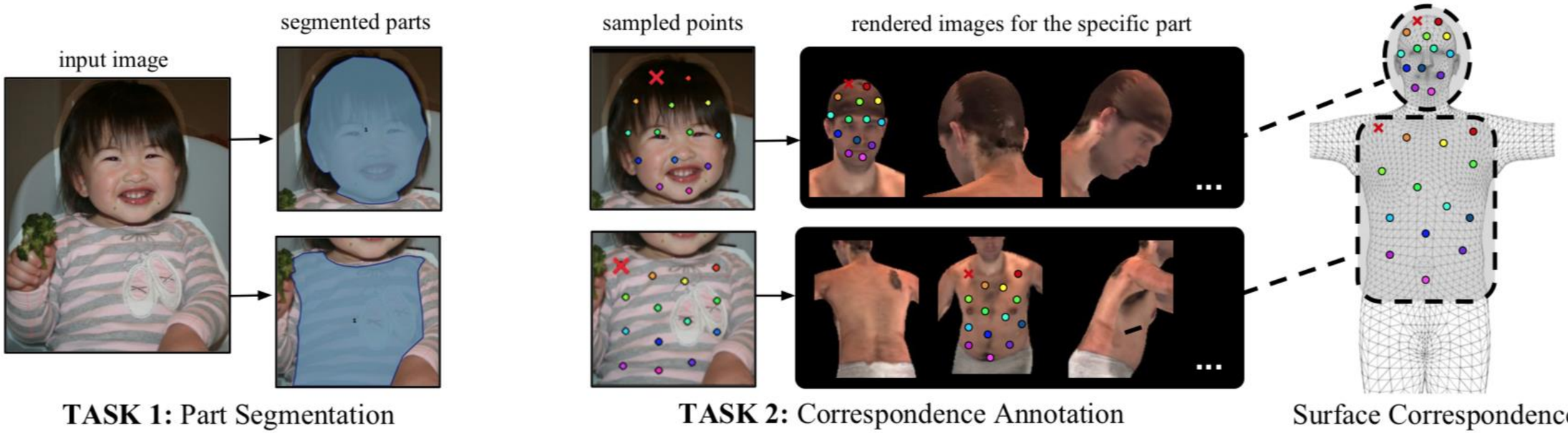
Natalia Neverova
Facebook AI Research

Iasonas Kokkinos
Facebook AI Research

* Rıza Alp Güler was with Facebook AI Research during this work.

First Step

Part Segmentation



TASK 1: Part Segmentation

TASK 2: Correspondence Annotation

Surface Correspondence

„We annotate dense correspondence between images and a 3D surface model by asking the annotators to first segment the image into semantic regions and then localize each of the sampled points on any of the rendered part images. The surface coordinates of the rendered views are used to localize the collected 2D points on the 3D model. „

Second Step

Coco Dataset

In this Section we introduce DensePose-COCO, a largescale dataset for dense human pose estimation. DensePose- COCO provides ground-truth for 50K humans and contains more than 5 million manually annotated pairs.

In the second stage we sample every part region with a set of roughly equidistant points obtained by running k-means over the coordinates occupied by each part and request the annotators to bring these points in correspondence with the surface.

Second Step

Accuracy of human annotators

In particular, we provide annotators with synthetic images generated through the rendering system and textures of [45]. We ask the annotators to bring the synthesized images into correspondence with the surface using our annotation tool, and for every image k estimate the geodesic distance di,k between the correct surface point, i and the point estimated by human annotators ˆik :

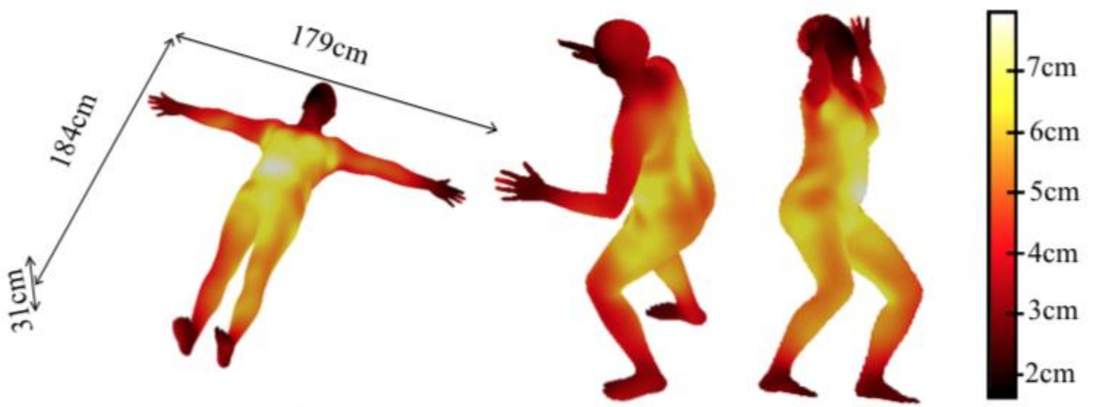
$$d_{i,k} = g(i, \hat{i}_k),$$

g(·, ·) is the geodesic distance between two surface points. For any image k, we annotate and estimate the error on a randomly sampled set of surface points Sk and interpolate the errors on the remainder of the surface. Finally, we average the errors across all examples given to the annotators.

[45] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2, 3, 6

Second Step

Per-instance evaluation



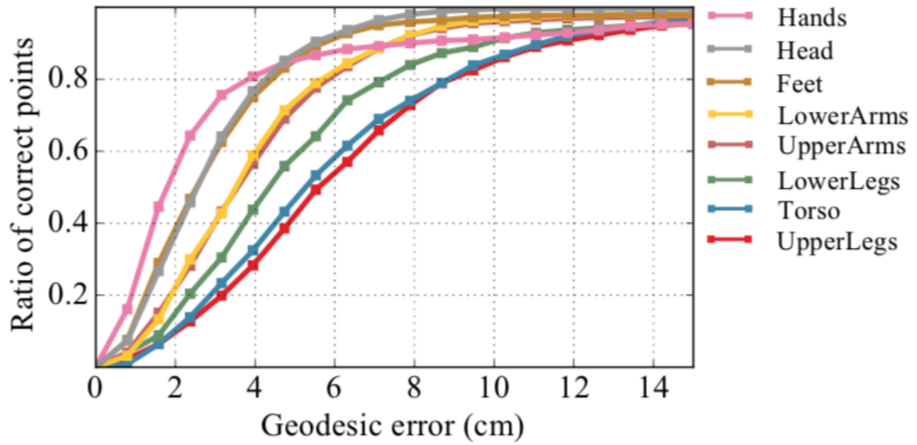
Inspired by the object keypoint similarity (OKS) measure used for pose evaluation on the COCO dataset , we introduce *geodesic point similarity (GPS)* as a correspondence matching score:

$$\mathrm{GPS}_j = \frac{1}{|P_j|} \sum_{p \in P_j} \exp\left(\frac{-g(i_p, \hat{i}_p)^2}{2\kappa^2} \right),$$

where P_j is the set of ground truth points annotated on person instance j, i_p is the vertex estimated by a model at point p, $\hat{i}_p$ is the ground truth vertex p and κ is a normalizing parameter. We set κ=0.255 so that a single point has a GPS value of 0.5 if its geodesic distance from the ground truth equals the average half-size of a body segment, cor- responding to approximately 30 cm. Intuitively, this means that a score of GPS ≈ 0.5 can be achieved by a perfect part segmentation model, while going above that also requires a more precise localization of a point on the surface.

Third Step

Fully-convolutional dense pose regression



We now turn to the task of training a deep network that predicts dense correspondences between image pixels and surface points. Such a task was recently addressed in the Dense Regression (DenseReg) system of through a fully-convolutional network architecture. In this Section we introduce improved architectures by combining the DenseReg approach with the Mask-RCNN architecture , yielding our ‘DensePose-RCNN’ system.

Using the surface representation, a simple choice for dense image-to-surface correspondence estimation consists in using a fully convolutional network (FCN) that combines a classification and a regression task, similar to DenseReg. In a first step, we classify a pixel as belonging to either background or one among the surface parts. In a second step, a regression system indicates the exact coordinates of the pixel within the part. Intuitively, we can say that we first use appearance to make a coarse estimate of where the pixel belongs to and then align it to the exact position through some small-scale correction.

Third Step

Fully-convolutional dense pose regression

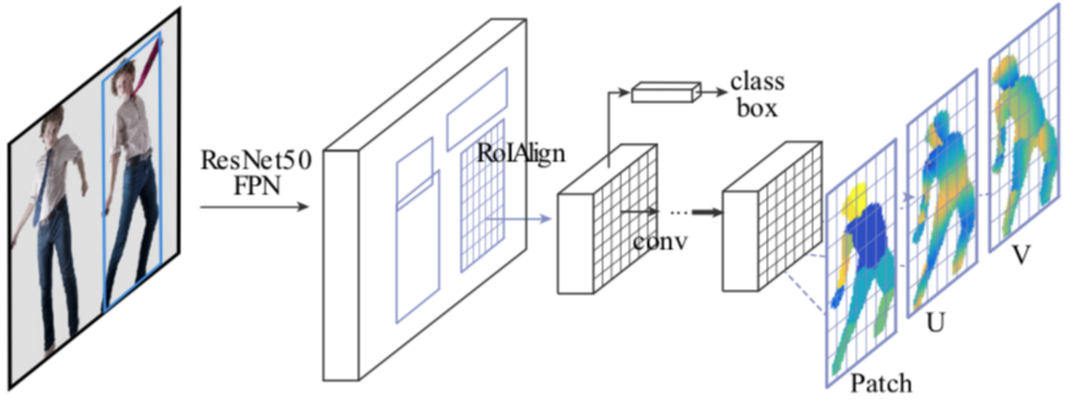


Figure 6: DensePose-RCNN architecture: we use a cascade of region proposal generation and feature pooling, followed by a fully-convolutional network that densely predicts discrete part labels and continuous surface coordinates.

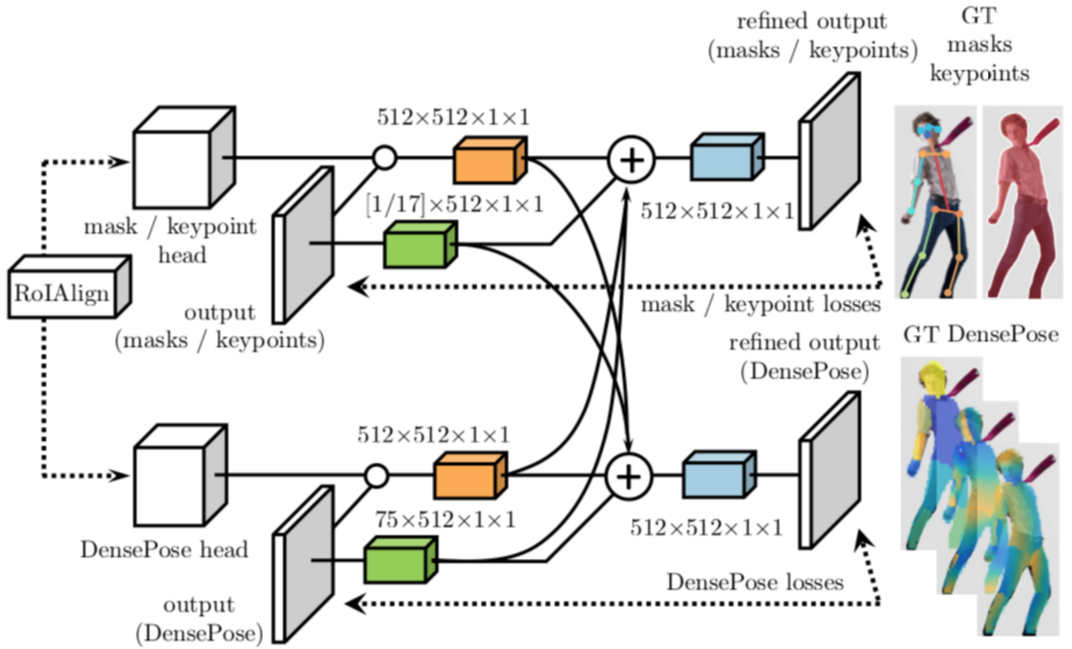


Figure 7: Cross-cascading architecture: The RoIAlign output in Fig. 6 feeds into the DensePose network and auxiliary networks for other tasks (masks, keypoints). Once first-stage predictions are obtained from all tasks, they are combined and fed into a second-stage refinement unit.

Thank you

Ionescu Vlad