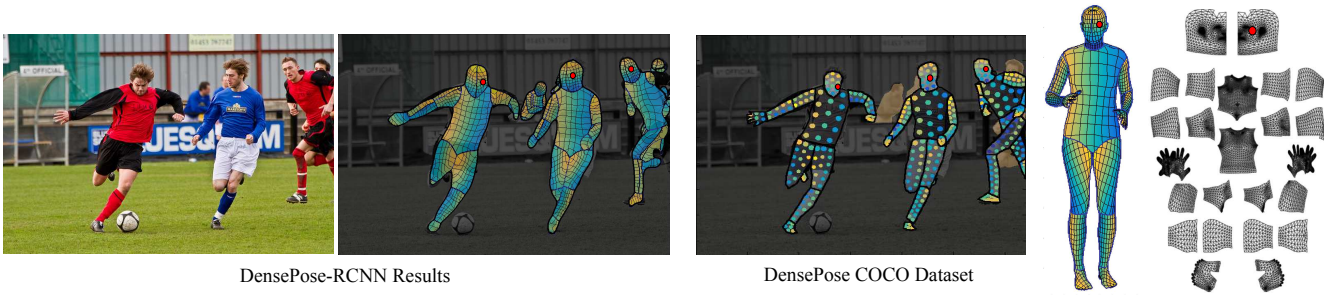
DensePose: Dense Human Pose Estimation In The Wild

Rıza Alp Güler*
INRIA-CentraleSupélec
riza.guler@inria.fr

Natalia Neverova
Facebook AI Research
nneverova@fb.com

Iasonas Kokkinos
Facebook AI Research
iasonask@fb.com



DensePose-RCNN Results | DensePose COCO Dataset

Figure 1: Dense pose estimation aims at mapping all human pixels of an RGB image to the 3D surface of the human body. We introduce DensePose-COCO, a large-scale ground-truth dataset containing manually annotated image-to-surface correspondences for 50K images, and train DensePose-RCNN to densely regress UV coordinates at multiple frames per second. *Left:* The image and the regressed correspondence by DensePose-RCNN. *Middle:* DensePose-COCO Dataset annotations. *Right:* Partitioning and UV parametrization of the body surface.

Abstract

In this work we establish dense correspondences between an RGB image and a surface-based representation of the human body, a task we refer to as dense human pose estimation. We gather dense correspondences for 50K persons appearing in the COCO dataset by introducing an efficient annotation pipeline. We then use our dataset to train CNN-based systems that deliver dense correspondence 'in the wild', namely in the presence of background, occlusions and scale variations. We improve our training set's effectiveness by training an inpainting network that can fill in missing ground truth values and report improvements with respect to the best results that would be achievable in the past. We experiment with fully-convolutional networks and region-based models and observe a superiority of the latter. We further improve accuracy through cascading, obtaining a system that delivers highly-accurate results at multiple frames per second on a single gpu. Supplementary materials, data, code, and videos are provided on the project page http://densepose.org.

1. Introduction

This work aims at pushing further the envelope of human understanding in images by establishing dense correspondences between a 2D image and a 3D, surface-based representation of the human body. We can understand this task as involving several other problems, such as object detection, pose estimation, part and instance segmentation either as special cases or prerequisites. Addressing this task has applications in problems that require going beyond plain landmark localization, such as graphics, augmented reality, or human-computer interaction, and could also be a stepping stone towards general 3D-based object understanding.

The task of establishing dense correspondences from an image to a surface-based model has been addressed mostly in the setting where a depth sensor is available [43, 34, 46]. Instead, we establish dense image-to-surface correspondences using as sole input the RGB values of a single image.

Several other works have recently aimed at recovering dense correspondences between pairs [3] or sets of RGB images [50, 10] in an unsupervised setting. More recently, [44] used the equivariance principle in order to align sets of images to a common coordinate system, while following the general idea of groupwise image alignment, e.g. [24, 22].

While these works target general categories, ours is focused on arguably the most important one, humans. For humans one can simplify the task by exploiting parametric deformable surface models, such as the Skinned Multi-Person Linear (SMPL) model [2], or the more recent Adam model [15] obtained through controlled 3D surface acquisition. Turning to the task of image-to-surface mapping,

[1] Rıza Alp Güler was with Facebook AI Research during this work.

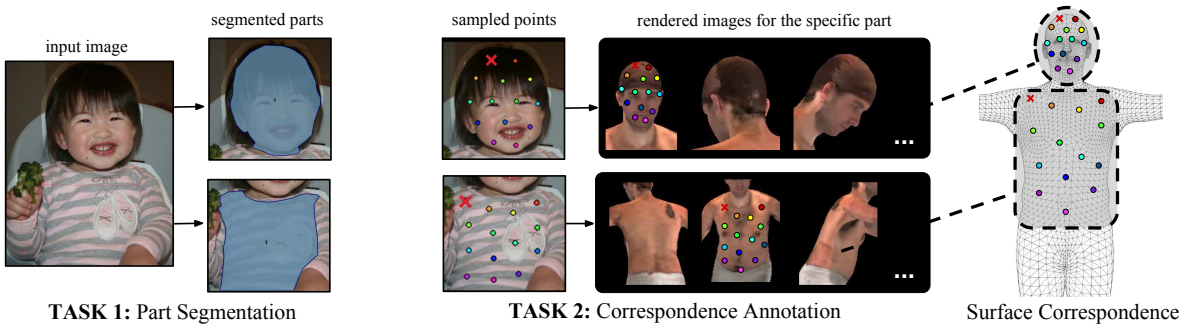


Figure 2: We annotate dense correspondence between images and a 3D surface model by asking the annotators to first segment the image into semantic regions and then localize each of the sampled points on any of the rendered part images. The surface coordinates of the rendered views are used to localize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

in [2], the authors propose a two-stage method of detecting human landmarks and fitting a parametric deformable surface model to the image through iterative minimization. In parallel to our work, [21] extend [2] to operate end-to-end, incorporating the iterative reprojection error minimization as a module of a deep network that recovers 3D camera pose and the low-dimensional body parametrization.

Our methodology differs from all these works in that we take a full-blown supervised learning approach and gather ground-truth correspondences between images and a detailed, accurate parametric surface model of the human body [28]: rather than using the SMPL model at test time we only use it as a means of defining our problem during training. Our approach can be understood as the next step in the line of works on human pose estimation [27, 1, 20, 7, 42, 19, 29]. Human part segmentation masks have been provided in a number of datasets [48, 6, 13]; these can be understood as providing a coarsened version of image-to-surface correspondence, where rather than continuous coordinates one predicts discretized part labels [35]. Surface-level supervision was only recently introduced for synthetic images in [45], while in [23] a dataset of 8515 images is annotated with keypoints and semi-automated fits of 3D models to images. In this work instead of compromising the extent and realism of our training set we introduce a novel annotation pipeline that allows us to gather ground-truth correspondences for 50K images of COCO, yielding our new DensePose-COCO dataset.

Our work is closest in spirit to the recent DenseReg framework [14], where CNNs were trained to establish dense correspondences between a 3D model and images 'in the wild'. That work focused mainly on faces, and provided evaluations on datasets with moderate pose variability. Here, however, we are facing new challenges, due to the higher complexity and flexibility of the human body, as well as the larger scale variation. We address these challenges by designing appropriate architectures (Sec. 3) that yield substantial improvements over a DenseReg-type fully convolutional architecture. By combining our approach with the recent Mask-RCNN system of [16] we show that a discriminatively trained model can efficiently recover highly-accurate correspondence fields for complex scenes involving tens of persons: on a GTX 1080 GPU our system operates at 20-26 fps for a 240×320 image or 4-5 fps for a 800×1100 image.

Our contributions can be summarized in three points. Firstly, as described in Sec. 2, we introduce the first manually-collected ground truth dataset for the task, by gathering dense correspondences between the SMPL model [28] and persons appearing in the COCO dataset. This is accomplished through a novel annotation pipeline that exploits 3D surface information during annotation.

Secondly, as described in Sec. 3, we use the resulting dataset to train CNN-based systems that deliver dense correspondence 'in the wild' by regressing body surface coordinates at any image pixel. We experiment with both fully-convolutional architectures, relying on Deeplab [4], and also with region-based systems, relying on Mask-RCNN [16], observing a superiority of the latter. We also consider cascading variants of our approach, yielding further improvements over existing architectures.

Thirdly, we explore different ways of exploiting our constructed ground truth information. Our supervision signal is defined over a randomly chosen subset of image pixels per training sample. We use these sparse correspondences to train a teacher network that can inpaint the supervision signal in the rest of the image domain. Using this inpainted signal results in better performance when compared to either sparse points, or any other existing dataset, as shown experimentally in Sec. 4.

Our experiments indicate that dense human pose estimation is to a large extent feasible, but still has space for improvement. Our code and the data will be made publicly available at http://densepose.org.

Figure 3: Visualization of annotations: Image (left), U (middle) and V (right) values for the collected points.

2. DensePose-COCO Dataset

Gathering rich, high-quality training sets has been a catalyst for progress in the classification [40], detection and segmentation [8, 27] tasks. There currently exists no manually collected ground-truth for dense human pose estimation for real images. The works of [23, 45] can be used as surrogates, but as we show in Sec. 4 provide worse supervision.

In this Section we introduce DensePose-COCO, a large-scale dataset for dense human pose estimation. DensePose-COCO provides ground-truth for 50K humans and contains more than 5 million manually annotated pairs. We first present our annotation pipeline, since our design choices may be useful for general 3D annotation. We then analyze the accuracy of the collected ground-truth, and finally introduce evaluation metrics for dense pose estimation.

2.1. Annotation System

In this work we use human annotators to establish dense correspondences from 2D images to surface-based representations of the human body. If done naively this would require manipulating a surface through rotations to find the vertices corresponding to every 2D image point, which is time-demanding and inefficient. Instead, we construct an annotation pipeline through which we can efficiently gather annotations for image-to-surface correspondence.

As shown in Fig. 2, in the first stage we ask annotators to delineate regions corresponding to visible, semantically defined body parts. These include Head, Torso, Lower/Upper Arms, Lower/Upper Legs, Hands and Feet. In order to simplify the UV parametrization we design the parts to be isomorphic to a plane, partitioning the upper and lower limbs and torso into frontal-back parts. For *head*, *hands* and *feet*, we use the manually obtained UV fields provided in the SMPL model [28]. For other parts we obtain the unwrapping via multi-dimensional scaling applied to pairwise geodesic distances. The UV fields for the resulting 24 parts are visualized in Fig. 1 (right).

We instruct the annotators to estimate the body part behind the clothes, so that for instance wearing a large skirt will not complicate the subsequent correspondence annotations. 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. The number of sampled points varies based on the size of the part and the maximum number of sampled points per part is 14. In order to simplify this task we 'unfold' the part surface by providing six pre-rendered views of the same body part and allow the user to place landmarks on any of them. This allows the annotator to choose the most convenient viewpoint by selecting one among six options instead of manually rotating the surface. As the user indicates a point on any of the rendered part views, its surface coordinates are used to simultaneously show its position on the remaining views – this gives a global overview of the correspondence. We show indicative visualizations of the gathered annotations in Fig. 3.

2.2. Accuracy of human annotators

A common concern when gathering ground-truth is the accuracy of the human annotations, which is often seen as an upper bound of what vision algorithms can deliver. In pose estimation one typically asks multiple annotators to label the same landmark, which is then used to assess the variance in position, e.g. [27, 38]. In our case we can directly compare to the true mesh coordinates used to render a pixel, rather than first estimating a 'consensus' landmark location among multiple 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 $d_{i,k}$ between the correct surface point, i and the point estimated by human annotators $\hat{i}_k$:

$$d_{i,k} = g(i, \hat{i}_k), \tag{1}$$

$g(\cdot, \cdot)$ 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 $\mathcal{S}_k$ and interpolate the errors on the remainder of the surface. Finally, we average the errors across all examples given to the annotators.

As shown in Fig. 4 the annotation errors are substantially smaller on small surface parts with distinctive features that could help localization (face, hands, feet), while on larger uniform areas that are typically covered by clothes (torso, back, hips) the annotator errors can get larger.

2.3. Evaluation Metrics

We consider two different ways of summarizing correspondence accuracy over the whole human body, including pointwise and per-instance evaluation.

Pointwise evaluation. This approach evaluates correspondence accuracy over the whole image domain through the Ratio of Correct Point (RCP) correspondences, where a correspondence is declared correct if the geodesic distance is below a certain threshold. As the threshold t varies, we obtain a curve $f(t)$, whose area provides us with a scalar summary of the correspondence accuracy. For any given image we have a varying set of points coming with ground-truth signals. We summarize performance on the ensemble of such points, gathered across images. We evaluate the area under the curve (AUC), $\text{AUC}_a = \frac{1}{a}\int_0^a f(t)dt$, for two different values of $a = 10\text{cm}, 30\text{cm}$ yielding AUC_{10} and AUC_{30} respectively, where AUC_{10} is understood as being an accuracy measure for more refined correspondence. This performance measure is easily applicable to both single- and multi-person scenarios and can deliver directly comparable values. In Fig. 5 we provide the per-part pointwise evaluation of the human annotator performance on synthetic data, which can be seen as an upper bound for the performance of our systems.

Per-instance evaluation. Inspired by the object keypoint similarity (OKS) measure used for pose evaluation on the COCO dataset [27, 38], we introduce *geodesic point similarity (GPS)* as a correspondence matching score:

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

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, corresponding 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.

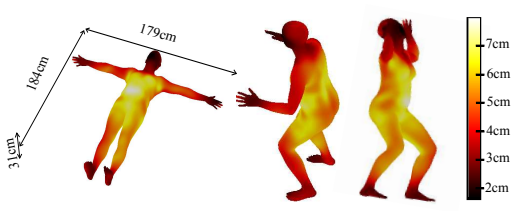


Figure 4: Average human annotation error on the surface.

Once the matching is performed, we follow the COCO challenge protocol [27, 39] and evaluate Average Precision (AP) and Average Recall (AR) at a number of GPS thresholds ranging from 0.5 to 0.95, which corresponds to the range of geodesic distances between 0 and 30 cm. We use the same range of distances to perform both per-instance and per-point evaluation.

3. Learning Dense Human Pose Estimation

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 [14] through a fully-convolutional network architecture [4]. In this Section we introduce improved architectures by combining the DenseReg approach with the Mask-RCNN architecture [16], yielding our 'DensePose-RCNN' system. We develop cascaded extensions of DensePose-RCNN that further improve accuracy and describe a training-based interpolation method that allows us to turn a sparse supervision signal into a denser and more effective variant.

3.1. Fully-convolutional dense pose regression

Since the human body has a complicated structure, we break it into multiple independent pieces and parametrize each piece using a local two-dimensional coordinate system, that identifies the position of any node on this surface part.

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.

Concretely, coordinate regression at an image position i

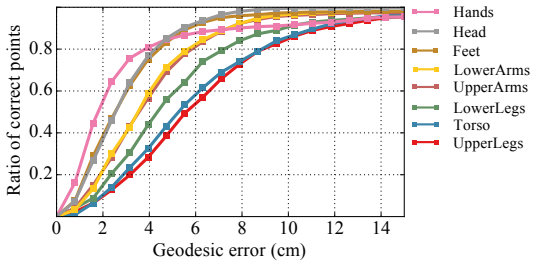


Figure 5: Human annotation error distribution within parts.

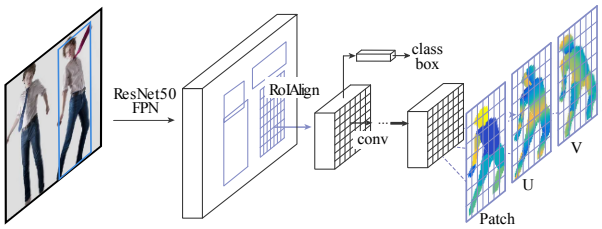


Figure 6: DensePose-RCNN architecture: we use a cascade of region proposal generation and feature pooling, followed by a fully-convolutional network that densely predicts discrete part labels and continuous surface coordinates.

can be formulated as follows:

$$c^* = \text{argmax}_c P(c|i), \quad [U, V] = R^{c^*}(i) \tag{3}$$

where in the first stage we assign position i to the body part c^* that has highest posterior probability, as calculated by the classification branch, and in the second stage we use the regressor R^{c^*} that places the point i in the continuous U, V coordinates parametrization of part c^*. In our case, c can take 25 values (one is background), meaning that P_x is a 25-way classification unit, and we train 24 regression functions R^c, each of which provides 2D coordinates within its respective part c. While training, we use a cross-entropy loss for part classification and a smooth L_1 loss [12] for each part-specific regression function. The regression loss for a part is only considered for pixels occupied by that part.

3.2. Region-based Dense Pose Regression

Using an FCN makes the system particularly easy to train, but loads the same deep network with too many tasks, including part segmentation and pixel localization, while at the same time requiring scale-invariance, which becomes challenging for humans in COCO. Here we adopt the region-based approach of [36, 16], which consists in a cascade of proposing regions-of-interest (ROI), extracting region-adapted features through ROI pooling [17, 16] and feeding the resulting features into a region-specific branch. Region-based architectures decompose the complexity of the task into controllable modules and implement a scale selection mechanism through ROI-pooling. At the same time, they can be jointly trained in an end-to-end manner [36].

We adopt the settings introduced in [16], involving the construction of Feature Pyramid Network [26] features, and ROI-Align pooling, which have been shown to be important for tasks that require spatial accuracy. We adapt this architecture to our task, so as to obtain dense part labels and coordinates within each of the selected regions.

As shown in Fig. 6, on top of ROI-pooling we introduce an FCN that is entirely devoted to these two tasks, generating a classification and a regression head that provide the part assignment and part coordinate predictions, as in

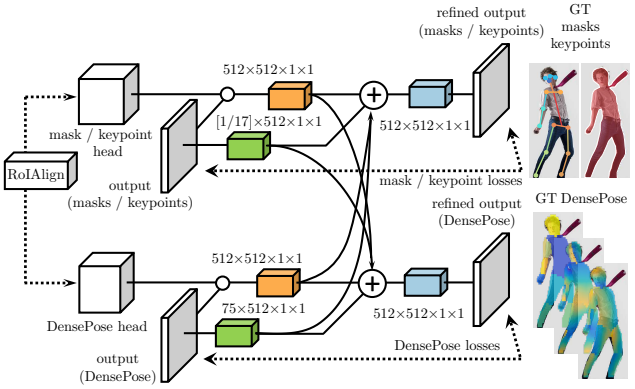


Figure 7: Cross-cascading architecture: The RoIAlign output in Fig. 6 feeds into the DensePose network and auxiliary networks for other tasks (masks, keypoints). Once first-stage predictions are obtained from all tasks, they are combined and fed into a second-stage refinement unit.

DenseReg. For simplicity, we use the exact same architecture used in the keypoint branch of Mask-RCNN, consisting of a stack of 8 alternating 3×3 fully convolutional and ReLU layers with 512 channels. At the top of this branch we have the same classification and regression losses as in the FCN baseline, but we now use a supervision signal that is cropped within the proposed region.

3.3. Multi-task cascaded architectures

Inspired by the success of recent pose estimation models based on iterative refinement [47, 31] we experiment with cascaded architectures. Cascading can improve performance both by providing context to the following stages, and also through the benefits of deep supervision [25].

As shown in Fig. 7 we do not confine ourselves to cascading within a single task, but also exploit information from related tasks, such as keypoint estimation and instance segmentation, which have successfully been addressed by the Mask-RCNN architecture [16]. This allows us to exploit task synergies and the complementary merits of different sources of supervision.

3.4. Distillation-based ground-truth interpolation

Even though we aim at dense pose estimation at test time, in every training sample we annotate only a sparse subset of the pixels, approximately 100-150 per human. This does not necessarily pose a problem during training, since we can make our classification/regression losses oblivious to points where the ground-truth correspondence was not collected, simply by not including them in the summation over the per-pixel losses [41]. However, we have observed that we obtain better results by "inpainting" the values of the supervision signal on positions that were not originally annotated. For this we adopt a learning-based ap-

proach where we firstly train a "teacher" network to reconstruct the ground-truth values wherever these are observed, and then deploy it on the full image domain, yielding a dense supervision signal.

As shown in Fig. 8, we use human segmentation maps available in COCO in order to get the most accurate supervision signal possible by (a) replacing background structures with a common gray value and (b) ignoring the network's predictions outside the human region. The performance of the teacher network can therefore be understood as an upper bound on what an algorithm can deliver on real data, since we remove false positives, normalize scale and remove background variation during both training and testing.

4. Experiments

In all experiments we assess the methods on a test set of 1.5K images containing 2.3K humans and use 48K humans in the training set. Our test set coincides with the COCO keypoints-minival partition used by [16] and the training set with the COCO-train partition.

Before assessing dense pose estimation *in the wild* (Sec. 4.2), we start in Sec. 4.1 with the 'Single-Person' setting where the input images are cropped around ground-truth boxes. This factors out the effects of detection performance and provides us with a controlled setting to assess the usefulness of the DensePose-COCO dataset.

4.1. Single-Person Dense Pose Estimation

In Sec. 4.1.1 we compare the DensePose-COCO dataset to other sources of supervision for dense pose estimation. In Sec. 4.1.2 we compare the performance of the model-based system of [2] with ours. We note that the system of [2] was not trained with the same amount of data as our model; this comparison therefore serves primarily to show the merit of our large-scale dataset for discriminative training.

4.1.1 Manual supervision versus surrogates

We start by assessing whether DensePose-COCO improves the accuracy of dense pose estimation with respect to the prior semi-automated, or synthetic supervision signals.

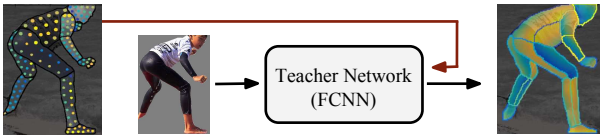


Figure 8: We train a 'teacher network' with our collected sparse supervision signal and use it to 'inpaint' a dense supervision signal used to train our region-based system.

A semi-automated method is used for the 'Unite the People' (UP) dataset of [23], where human annotators verified the results of fitting the SMPL 3D deformable model [28] to 2D images. However, model fitting often fails in the presence of occlusions, or extreme poses, and is never guaranteed to be entirely successful – for instance, even after rejecting a large fraction of the fitting results, the feet are still often misaligned in [23].

Synthetic ground-truth can be established by rendering images using surface-based models [33, 32, 37, 11, 5, 30]. This has recently been applied to human pose in the SURREAL dataset of [45], where the SMPL model [28] was rendered with the CMU Mocap dataset poses [29]. However, domain shift can emerge because of the different statistics of rendered and natural images.

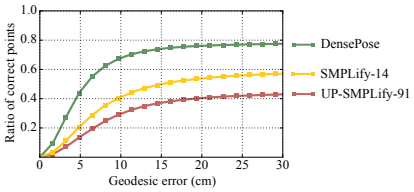
Since both of these two methods use the same SMPL surface model as the one we use in our work, we can directly compare results, and also combine datasets. We render our dense coordinates and our dense part labels on the SMPL model for all 8514 images of UP dataset and 60k SURREAL models for comparison.

In Fig. 10 we assess the test performance of ResNet-101 FCNs of stride 8 trained with different datasets, using a Deeplab-type architecture. During training we augment samples from all of the datasets with scaling, cropping and rotation. We observe that the surrogate datasets lead to weaker performance, while their combination yields improved results. Still, their performance is substantially lower than the one obtained by training on our DensePose dataset, while combining the DensePose with SURREAL results in a moderate drop in network performance. Based on these results we rely exclusively on the DensePose dataset for training in the remaining experiments, even though domain adaptation [9] could be used in the future to exploit synthetic sources of supervision.

The last line in the table of Fig. 10 ('DensePose*') indicates the additional performance boost that we get by using the teacher network settings described in Sec. 3.4. Clearly, the results are not directly comparable with those of other methods, since we use additional information to remove background structures. Still, the resulting predictions are substantially closer to human performance – we can therefore confidently use our teacher network to obtain dense supervision for the experiments in Sec. 4.2.

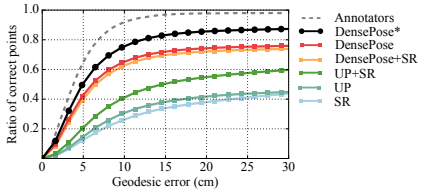
4.1.2 FCNN- vs Model-based pose estimation

In Fig. 9 we compare our method to the SMPLify pipeline of [2], which fits the 3D SMPL model to an image based on a pre-computed set of landmark points. We use the code provided by [23] with both DeeperCut pose estimation landmark detector [18] for 14-landmark results and with the 91-landmark alternative proposed in [23].



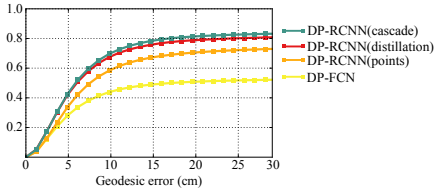
Method	AUC_{10}	AUC_{30}
Full-body images		
UP-SMPLify-91	0.155	0.306
SMPLify-14	0.226	0.416
DensePose	0.429	0.630
All images		
SMPLify-14	0.099	0.19
DensePose	0.378	0.614
Human Performance	0.563	0.835

Figure 9: Qualitative comparison between model-based single-person pose estimation of SMPLify [2] and our FCN-based result, in the absence ('full-body images') and presence ('all images') of occlusions.



Method	AUC_{10}	AUC_{30}
SR	0.124	0.289
UP	0.146	0.319
SR + UP	0.201	0.424
DensePose + SR	0.357	0.592
DensePose	0.378	0.614
DensePose*	0.445	0.711
Human Performance	0.563	0.835

Figure 10: Single-person performance for different kinds of supervision signals used for training: DensePose leads to substantially more accurate results than surrogate datasets. DensePose* uses a figure-ground oracle at both training and test time.



Method	AUC_{10}	AUC_{30}	IoU
DP-FCN	0.253	0.418	0.66
DP-RCNN (points only)	0.315	0.567	0.75
DP-RCNN (distillations)	0.381	0.645	0.79
DP-RCNN (cascade)	0.390	0.664	0.81
DP*	0.417	0.683	–
Human Performance	0.563	0.835	–

Figure 11: Results of multi-person dense correspondence labelling. Here we compare the performance of our proposed DensePose-RCNN system against the fully-convolutional alternative on realistic images from the COCO dataset including multiple persons with high variability in scales, poses and backgrounds.

Since the whole body is visible in the MPII dataset used for training the landmark detectors, for a fair comparison we separately evaluate on images where 16/17 or 17/17 landmarks are visible and on the whole test set. We observe that while being orders of magnitude faster (0.04-0.25" vs 60-200") our bottom-up method largely outperforms the iterative, model fitting result. As mentioned above, this difference in accuracy indicates the merit of having at our disposal DensePose-COCO for discriminative training.

4.2. Multi-Person Dense Pose Estimation

Having established the merit of the DensePose-COCO dataset, we now turn to examining the impact of network architecture on dense pose estimation in-the-wild. In Fig. 11 we summarize our experimental findings using the same RCP measure used in Fig. 10.

We observe firstly that the FCN-based performance in-the-wild (curve 'DensePose-FCN') is now substantially lower than that of the DensePose curve in Fig. 11. Even though we apply a multi-scale testing strategy that fuses probabilities from multiple runs using input images of different scale [49], the FCN is not sufficiently robust to deal with the variability in object scale.

We then observe in curve 'DensePose-RCNN' a big boost in performance thanks to switching to a region-based system. The networks up to here have been trained using the sparse set of points that have been manually annotated. In curve 'DensePose-RCNN-Distillation' we see that using the dense supervision signal delivered by our DensePose* system on the training set yields a substantial improvement. Finally, in 'DensePose-RCNN-Cascade' we show the performance achieved thanks to the introduction of cascading: Sec. 3.3 almost matches the 'DensePose*' curve of Fig. 10.

This is a remarkably positive result: as described in Sec. 3.4, the 'DensePose*' curve corresponds to a very privileged evaluation and can be understood as an upper bound of what one can expect to obtain when operating in-the-wild. We see that our best system is marginally below that level of performance, which clearly reveals the power of the three modifications we introduce, namely region-based processing, inpainting the supervision signal, and cascading.

In Table 1 we report the AP and AR metrics described in Sec. 2 as we change different choices in our architecture. We have conducted experiments using both ResNet-50 and ResNet-101 backbones and observed an only insignificant boost in performance with the larger model (first two rows in Table 1). The rest of our experiments are therefore based on the ResNet-50-FPN version of DensePose-RCNN. The following two experiments shown in the middle section of Table 1 indicate the impact on multi-task learning.

Augmenting the network with the mask or keypoint branches yields improvements with any of these two auxiliary tasks. The last section of Table 1 reports improvements in dense pose estimation obtained through the cascading setup from Fig. 7. Incorporating additional guidance in particular from the keypoint branch significantly boosts performance.

Our qualitative results in Fig. 12 indicate that our method is able to handle large amounts of occlusion, scale, and pose variation, regardless of the shape of the clothes.

Figure 12: Qualitative evaluation of DensePose-RCNN. *Left:* input, *Right:* DensePose-RCNN estimates. Our system successfully estimates body pose regardless of skirts or dresses, while handling a large variability of scales, poses, and occlusions.

Method	**AP**	**AP**$_{50}$	**AP**$_{75}$	**AP**$_M$	**AP**$_L$	**AR**	**AR**$_{50}$	**AP**$_{75}$	**AR**$_M$	**AR**$_L$
DensePose (ResNet-50)	51.0	83.5	54.2	39.4	53.1	60.1	88.5	64.5	42.0	61.3
DensePose (ResNet-101)	51.8	83.7	56.3	42.2	53.8	61.1	88.9	66.4	45.3	62.1
Multi-task learning										
DensePose + masks	51.9	85.5	54.7	39.4	53.9	61.1	89.7	65.5	42.0	62.4
DensePose + keypoints	52.8	85.6	56.2	42.2	54.7	62.6	89.8	67.7	45.4	63.7
Multi-task learning with cascading										
DensePose-ST	51.6	83.9	55.2	41.9	53.4	60.4	88.9	65.3	43.3	61.6
DensePose + masks	52.8	85.5	56.1	40.3	54.6	62.0	89.7	67.0	42.4	63.3
DensePose + keypoints	55.8	87.5	61.2	48.4	57.1	63.9	91.0	69.7	50.3	64.8

Table 1: Per-instance evaluation of DensePose-RCNN performance on COCO `minival`. All multi-task experiments are based on ResNet-50. DensePose-ST applies cascading to the base, single-task network.

5. Conclusion

In this work we have addressed the task of dense human pose estimation using discriminatively trained models. We introduce DensePose-COCO, a large-scale dataset of ground-truth image-surface correspondences and develop novel architectures for recovering highly-accurate dense correspondences between images and the body surface in multiple frames per second. We anticipate that this will lead to novel augmented reality or graphics tasks, and we intend to further pursue the association of images with semantic 3D object representations.

Acknowledgements

We thank the authors of [16] for their code, P. Dollar and T.-Y. Lin for help with COCO, the authors of [28] for making the SMPL model open for research and H. Y. Güler for his help with back-end development.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 2

[2] F. Bogo, A. Kanazawa, C. Lassner, P. V. Gehler, J. Romero, and M. J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 1, 2, 6, 7

[3] H. Bristow, J. Valmadre, and S. Lucey. Dense semantic correspondence where every pixel is a classifier. In *ICCV*, 2015. 1

[4] L. 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2, 4

[5] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *3DV*, 2016. 6

[6] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 2

[7] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Body parts dependent joint regressors for human pose estimation in still images. *IEEE Transactions on pattern analysis and machine intelligence*, 36(11):2131–2143, 2014. 2

[8] M. Everingham, S. M. A. Eslami, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 3

[9] Y. Ganin and V. S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015. 6

[10] U. Gaur and B. S. Manjunath. Weakly supervised manifold learning for dense semantic object correspondence. In *ICCV*, 2017. 1

[11] M. F. Ghezelghieh, R. Kasturi, and S. Sarkar. Learning camera viewpoint using cnn to improve 3d body pose estimation. In *3DV*, 2016. 6

[12] R. B. Girshick. Fast R-CNN. In *ICCV*, 2015. 5

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017. 2

[14] R. A. Güler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *CVPR*, 2017. 2, 4

[15] Y. S. Hanbyul Joo, Tomas Simon. Total capture: A 3d deformation model for tracking faces, hands, and bodies. *arXiv preprint arXiv:1801.01615*, 2018. 1

[16] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask r-cnn. *CVPR*, 2017. 2, 4, 5, 6, 8

[17] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014. 5

[18] E. Insafutdinov, L. Pishchulin, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. 6

[19] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2

[20] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2

[21] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 2

[22] I. Kokkinos and A. L. Yuille. Unsupervised learning of object deformation models. In *ICCV*, 2007. 1

[23] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 2, 3, 6

[24] E. G. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):236–250, 2006. 1

[25] C. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *AISTATS*, 2015. 5

[26] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 5

[27] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 2, 3, 4

[28] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. 2, 3, 6, 8

[29] mocap.cs.cmu.edu. Cmu graphics lab motion capture database, 2003. 2, 6

[30] N. Neverova, C. Wolf, F. Nebout, and G. Taylor. Hand pose estimation through weakly-supervised learning of a rich intermediate representation. *Computer Vision and Image Understanding*, 2017. 6

[31] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 5

[32] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, 2012. 6

[33] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele. Learning people detection models from few training samples. In *CVPR*, 2011. 6

[34] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon. Metric regression forests for correspondence estimation. *International Journal of Computer Vision*, 113(3):163–175, 2015. 1

[35] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *CVPR*, 2017. 2

[36] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 5

[37] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *NIPS*, 2016. 6

[38] M. R. Ronchi and P. Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *ICCV*, 2017. 3, 4

[39] M. R. Ronchi and P. Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *ICCV*, 2017. 4

[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 3

[41] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017. 5

[42] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010. 2

[43] J. Taylor, J. Shotton, T. Sharp, and A. W. Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *CVPR*, 2012. 1

[44] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised object learning from dense equivariant image labelling. In *NIPS*, 2017. 1

[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

[46] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li. Dense human body correspondences using convolutional networks. In *CVPR*, 2016. 1

[47] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 5

[48] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 2

[49] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 7

[50] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *CVPR*, 2016. 1