28^{th} *Computer Vision Winter Workshop*
Graz, Austria, February 12–14, 2025

DOI: 10.3217/978-3-99161-022-9-003

Human Pose-Constrained UV Map Estimation

Matej Suchanek, Miroslav Purkrabek, Jiri Matas

Visual Recognition Group
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague

sucham11@fel.cvut.cz

Abstract

UV map estimation is used in computer vision for detailed analysis of human posture or activity. Previous methods assign pixels to body model vertices by comparing pixel descriptors independently, without enforcing global coherence or plausibility in the UV map. We propose Pose-Constrained Continuous Surface Embeddings (PC-CSE), which integrates estimated 2D human pose into the pixel-to-vertex assignment process. The pose provides global anatomical constraints, ensuring that UV maps remain coherent while preserving local precision. Evaluation on DensePose COCO demonstrates consistent improvement, regardless of the chosen 2D human pose model. Whole-body poses offer better constraints by incorporating additional details about the hands and feet. Conditioning UV maps with human pose reduces invalid mappings and enhances anatomical plausibility. In addition, we highlight inconsistencies in the ground-truth annotations.

1. Introduction

Analysis of human pose is an essential part of many computer vision problems and is used in a number of applications, including recognition of human activity, gestures and interaction, detection of people and their intent in autonomous driving scenarios, etc.

Information about the human body can be estimated at different levels of resolution. The simplest is the detection of a bounding box that surrounds the person depicted. This can be more precisely delineated by body segmentation. *Pose estimation*, which estimates the locations of some body keypoints, provides another level of granularity. The most detailed is provided by *UV map estimation* (UVME), where every image pixel is mapped to the surface of a generalized human body. The surface is represented as a mesh with a fixed set of vertices.

The state-of-the-art methods for these tasks [10, 20, 27] rely on supervised learning, which possibly requires a large amount of annotated data. The cost and effort to annotate the data for human detection, segmentation, pose estimation, and UV map estimation increases with the complexity of the underlying task. UVME is arguably the most complex of these tasks and, therefore, the most data-hungry.

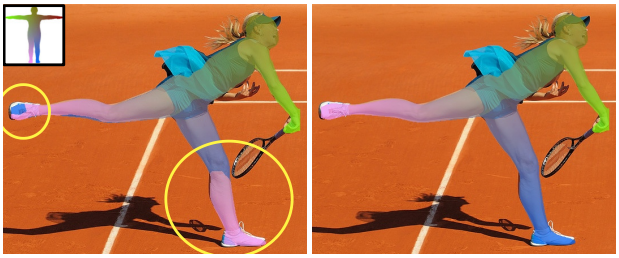


Figure 1. The Continuous Surface Embedding method (CSE) [20] (left) vs. Pose-Constrained CSE (right). The CSE method assigns each pixel of body segmentation to a vertex, and thus UV coordinate, on a canonical body shape mesh. The CSE assigns each pixel independently, leading to artifacts such as limb duplication (yellow circles). PC-CSE uses pose constraints during UV map estimation, producing smoother maps and eliminating artifacts. The UV values at individual pixels are visualized by color coding. The location of a given color on the canonical surface is shown in the inset image at the top left.

In a recent paper, a method for UVME called Continuous Surface Embeddings (CSE) was introduced [20]. The accuracy of the method is good, but it also has limitations. Due to the disparity between the resolution of the input image and the relatively small number of vertices, this method cannot perform one-to-one matching. Since each pixel is mapped independently of the others, the method can assign the same body part to multiple locations in the image or produce undesirable artifacts. Examples can be seen in Fig. 1 and 3.

In this paper, our objective is to leverage the methods for pose estimation, which have been in development for a considerable amount of time, to make UVME more accurate. We take advantage of their robustness and design, which guarantees no duplicate assignments. We introduce the concept of *pose-induced proximal regions* which constrain the mapping to a particular body part and propagate these constraints to the corresponding pixels.

We present a novel method called Pose-Constrained CSE (PC-CSE) that demonstrates the effectiveness of these concepts. It makes UV maps more coherent with essentially no loss of efficiency besides the need to calculate the human pose. PC-CSE shows consistent improvement over unconstrained UV maps. We conducted a detailed ablation study to justify our design choices and explain the improvement in performance.

2. Related Work

Human Pose Estimation (HPE) and UV Map Estimation (UVME) are closely related tasks. UVME provides more detailed and comprehensive information, while HPE benefits from a longer history of research, larger datasets, and greater robustness. In this work, we condition UVME predictions on HPE due to HPE's superior reliability. To establish context, we first discuss related work on HPE before moving to UVME advancements.

Data. Progress in human pose and gesture understanding relies heavily on large-scale datasets. The COCO dataset [16], with over 200,000 annotated images of people, is the most widely used, supporting tasks like object detection, instance segmentation, and pose estimation. Its annotations have been extended to whole-body keypoints [12] and UV map annotations [8]. Other datasets, such as MPII [3], CrowdPose [15], and OCHuman [28], target specific challenges like crowded scenes or people in close proximity. While these datasets have significantly advanced research, there is limited research on their overall annotation quality [22].

Current **2D Human Pose Estimation (HPE)** methods are categorized into top-down, bottom-up, and hybrid approaches. Top-down methods [18, 23, 27] first detect individuals using off-the-shelf person detectors, followed by pose estimation for each detected instance. ViTPose [27] represents the state-of-the-art in this category. Bottom-up methods [4, 6, 21] predict all keypoints simultaneously and group them into individual poses, making them more effective in crowded scenarios, such as those encountered in OCHuman [28]. Hybrid approaches [29] combine elements of both strategies, striking a balance between accuracy and efficiency under challenging conditions.

UV Map Estimation (UVME) has seen steady progress in recent years. DenseReg [7] formulates UVME as a regression task and trains a fully convolutional neural network for human face extraction using facial landmarks. DensePose [8], a milestone in UVME, collects a dataset of many body-to-surface annotations and adapts the Mask R-CNN architecture [9] for person detection, segmentation and UV map estimation in a cascade. Subsequent works focus on seeking correspondences in sequences of images [19, 24], utilize DensePose as an intermediate representation for other advanced tasks, such as 3D body reconstruction [2, 14], or use it as the ground truth [11].

DensePose relies on splitting the body template into small partitions ("charts") and performs a simultaneous regression of the target body part and the UV coordinate within the respective partition. Continuous Surface Embeddings (CSE) [20] follows up on DensePose by eliminating the need for artificial slicing of the template. Instead, CSE holds trainable descriptors (embeddings) of the template surface and guides a neural network to regress these embeddings per pixel in a contrastive manner. The UV map is determined by finding the closest surface embedding of every pixel. Overall, CSE simplifies the DensePose framework while making it generalizable to other natural objects. Both DensePose and CSE are tightly bound to the mesh of the SMPL [17], a parametrized 3D model of the human body.

BodyMap [10] further refines CSE by addressing body details such as hair and clothing, providing high-fidelity results while relying on CSE descriptors internally. Although it claims state-of-the-art performance, its code has not been released to the public. Recently, foundational models like Sapiens [13] have emerged in human-centric vision tasks. Trained on vast amounts of unannotated data, these models achieve state-of-the-art performance across various downstream tasks. However, they are resource-intensive and have yet to demonstrate significant advancements, specifically in UV map estimation.

3. Method

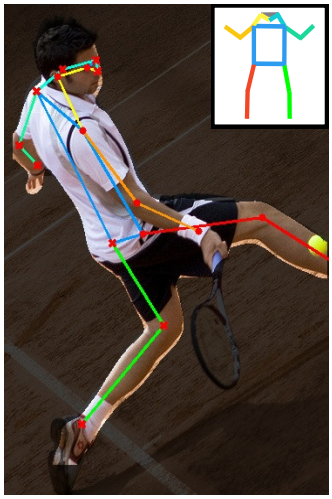
Our method is built on top of the CSE method [20], a feed-forward neural network based on the Mask R-CNN architecture [9]. Although it performs human detection, segmentation and UV map estimation in a cascade, we are concerned only with the latter and consider bounding boxes and segmentation as input determined by an external method.

The network outputs pixel descriptors, or *pixel embeddings*. During training, contrastive learning is employed to determine both the best weights and the values of *vertex embeddings*, each linked to one of the vertices of the SMPL mesh [17]. The resulting UV map is established by mapping every input pixel embedding to the most similar vertex embedding (in terms of cosine similarity), associating every image pixel with a mesh vertex (and its UV coordinates).

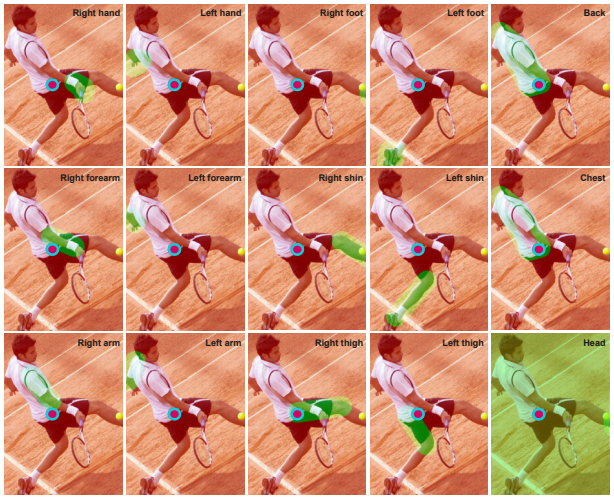
Formally, let I be the input image, $x \in I$ a (foreground) image pixel, $\Phi_x(I) \in \mathbb{R}^D$ embedding of the pixel x provided by the neural network Φ (where D is the embedding dimensionality), M the mesh (set of vertices), $i \in M$ a vertex index, and $E_i \in \mathbb{R}^D$ normalized embedding of the vertex i. The mapping from pixels to vertices using CSE [20] can be expressed as:

$$i_x^* = \arg\max_{i \in M} \langle E_i, \Phi_x(I) \rangle \,. \tag{1}$$

Consistent with the standard definition of mapping, CSE always maps exactly one vertex to every foreground pixel. However, the reverse is not necessarily true. Typically, the resolution of the input image is sufficiently high such that the pixel count significantly surpasses the vertex count on the mesh, resulting in multiple pixels being



(a) PC-CSE requires a bounding box and a segmentation mask as an input. VitPose-l [27] is used for pose estimation in this example. Front-view skeleton is in the inset image.



(b) Proximal regions of body parts. Only pixels in the segmentation mask and the green areas may be assigned to the body parts denoted in the top right. Head and undetected body parts are unconstrained (bottom right). The highlighted point can be assigned to the back, chest, head, and both the left and the right thigh but not to other parts.

(c) The PC-CSE UV map is consistent with the estimated pose, unlike the CSE [20] estimate. The difference is shown in Fig. 3.

Figure 2. Pose-constrained CSE (PC-CSE) takes an estimated bounding box, segmentation mask, and 2D human pose (a) as input. It computes proximal regions (b) for each body part and assigns pixels to SMPL [17] vertices to generate a UV map. Unlike the CSE [20], PC-CSE constrains pixel assignments using proximal regions, ensuring the resulting UV map aligns with the estimated pose (c).

frequently associated with the same vertex.

Arguably, this does not pose a problem in itself. For example, it is acceptable for neighboring pixels to map to the same vertex, as they can lie so close to each other on the actual body that the discretization of the mesh cannot distinguish between them. Nevertheless, a fully independent assignment of vertices to foreground pixels makes CSE generally prone to implausible pose predictions, as it has no other means to avoid them but to rely on the strength of its prior. Qualitative research confirms our hypothesis, as we observe situations stemming from the general problem, such as CSE assigning the same body part to more than one image region (*e.g.*, two hands are declared left), UV map discontinuities and various artifacts (see Fig. 3).

At this point, we examine the features of human pose estimation (HPE) algorithms. These estimators predict the locations of various landmarks on the human body, called *keypoints*, such as skeletal joints or facial landmarks. In particular, skeletal joints form a primitive human skeleton, the shape of which is very similar to that of our 3D human representation (Fig. 2a). Furthermore, each keypoint is, by design, assigned to at most one image coordinate. This constitutes the key advantage of HPE over CSE, as duplicate assignments of body parts become impossible.

3.1. Conditioning CSE by pose

We believe that using a human pose estimation model as a secondary expert during inference and enforcing consistency of the two representations is a promising path for avoiding errors in predicted UV maps and improving their quality. Therefore, we propose our new method called **Pose-Constrained Continuous Surface Embeddings** (PC-CSE). The key enhancement is the introduction of *pose-induced constraints* whose purpose is to limit the mapping of every pixel to only pre-selected body regions. It does not involve any architectural change to CSE and does not require its retraining or fine-tuning.

The constraints are rules that determine to which vertices of the mesh each foreground pixel is allowed to map. Which pixels are constrained by which rule depends on the inferred pose. We first define the relation between the pose representation and the target mesh. We use the COCO skeleton [16] as the default pose representation. It consists of 17 keypoints (Fig. 2a): 12 skeletal joints (wrists, elbows, shoulders, hips, knees, and ankles in pairs) and 5 facial landmarks (eyes, ears, and nose). These keypoints can be linked into arms, forearms, thighs, shins, and a quadrilateral defined by shoulders and hips. We refer to these connections as the *principal bones*.

In addition, we explore the *whole-body* skeleton [12]. This representation with 133 keypoints extends the COCO skeleton by introducing extra keypoints for hands, feet, and face. This poses an advantage over the basic version because hands and feet are somewhat distant from the respective keypoints and can deviate from the limb axis.

The canonical mesh can now be partitioned into subsets of vertices. Each partition should roughly correspond to one principal bone. We create 15 mesh partitions of SMPL – arms, forearms, hands, thighs, shins, and feet in pairs, the front and back of the torso, and head – by merging segments of SMPL body segmentation [1, 17]. We divide the torso by the sagittal plane to distinguish between the front and back of it.

The scope of constraints within the image is specified

by expanding ("inflating") the inferred skeleton composed of the principal bones. Each principal bone delineates its *proximal region*, each defined as a set of pixels with a certain maximum pixel distance from the bone (Fig. 2b). The optimal distance obviously relies on the apparent size of the person (which varies with its distance from the camera) and needs to be determined for each person separately. We try to estimate it using an algorithm that also depends on the pose; it is described in detail in Sec. 3.2.

The capsular shape of the proximal regions is most appropriate for the limbs, *i.e.*, arms, forearms, thighs, and shins. Concerning the front and back of the torso, we first merge the central quadrilateral (*i.e.*, between the shoulders and hips) with the regions around its sides, which we also define as having a capsule-like shape. Then, we analyze the mutual position of its corners to discriminate between the frontal and dorsal view. If the orientation of the keypoints implies the frontal view of the person, we subtract the quadrilateral from the back, and vice versa (see the rightmost column of Fig. 2b).

Nonetheless, the basic COCO skeleton does not adequately support precise localization of the hands (fingers) and feet (toes). Various strategies can be employed to manage this. With the whole-body skeleton, the proximal regions for these body parts can span the extra keypoints. As a fallback when using the basic skeleton, we propose circular proximal regions around the closest keypoint (wrist for hands, ankle for feet) twice as wide as the capsular ones. Both these options are discussed in the experiments (Sec. 5). Otherwise, a conservative approach is to merge body parts with the nearest bone or leave them unconstrained, but this does not fully leverage the capabilities of our method. In addition, we do not outline a dedicated proximal region for the head, but we let all pixels map to it. We consider the head to be easily recognizable, and our primary goal is to resolve duplications between paired limbs.

The proximal regions induce semantic labeling of image pixels by template partitions. Every pixel is labeled according to the proximal regions to which it belongs. If multiple proximal regions overlap, the pixels within the intersection are labeled with all corresponding labels. If a pixel falls outside all proximal regions, it gets all possible labels (thus, it keeps the original prediction). When a body part is missing (that is, either of its keypoints is not provided by the HPE model), we allow mapping to it from any foreground pixel. The purpose of this rule is to prevent inaccurate refinements where, for example, a forearm is partially visible, but one of its ends lies outside the image. As described earlier, we always apply this rule to the head as well.

As a result, each pixel receives information about its target body part(s) implied by the pose-induced constraints and the embedding provided by the original CSE. We now modify the original procedure (Eq. (1)) to consider the constraints as well. Instead of yielding the vertex with the highest similarity of all mesh vertices, we limit the output space to one of those vertices that belong to the body partitions defined by the constraints. The chosen vertex (its embedding) should still have the highest similarity to the pixel embedding, but only vertices from the limited subset of the whole mesh should be considered.

Formally, let $p \in P$ be the partition label (index), $M_p \subset M$ the vertices of the partition p, $L\colon I \to \mathcal{P}(P) \setminus \{\emptyset\}$ a function mapping a pixel to a set of allowed partitions. Equation (1) now becomes:

$$i_x^* = \arg \max_{i \in M_{L(x)}} \langle E_i, \Phi_x(I) \rangle \,, \tag{2}$$

where

$$M_{L(x)} = \bigcup_{p \in L(x)} M_p. \tag{3}$$

Alternatively, let $B(x,p)$ be the binary flag (0 or 1) indicating whether partition p is allowed in pixel x, $V(x,p)$ the vertex from partition p with the highest similarity to pixel x, $S(x,p)$ the similarity of vertex $V(x,p)$ to pixel x and $S'(x,p)$ our adjusted similarity. We compute these matrices as follows:

$$B(x,p) = [\![p \in L(x)]\!], \tag{4}$$

$$V(x,p) = \arg \max_{i \in M_p} \langle E_i, \Phi_x(I) \rangle \,, \tag{5}$$

$$S(x,p) = \max_{i \in M_p} \langle E_i, \Phi_x(I) \rangle \,, \tag{6}$$

$$S' = S \odot B. \tag{7}$$

Equation (2) is then equivalent to:

$$i_x^* = V(x, \arg \max_{p \in P} S'(x,p)). \tag{8}$$

We believe that this approach is more practical for implementation as it avoids computing unions of mesh partitions (Eq. (3)) and storing them in memory.

3.2. Determining proximal regions

As an intermediate step, PC-CSE expands the inferred human skeleton so that its shape approximately matches the silhouette (segmentation) of the person. The exact expansion range is a trade-off. Small proximal regions might not adjust the UV map at full width. Large proximal regions can cause significant overlaps with each other, making pose-induced constraints less effective. In extreme cases, the expansion range can be chosen as zero, resulting in no correction made, or it can be chosen so high that every proximal region covers the whole body. We note that in both cases, the new prediction would be the same as, and thus *not worse than*, the original prediction.

The expansion range should roughly correspond to the width (thickness) of the person's limbs, expressed in pixel units. We further refer to it as the *bone width* (Δ) and assume that it is proportional to other measures of the body, in particular the person's height. Typically, information about body measures is accessible only in controlled environments where the camera model and relative location of

the object and camera are known. However, this requirement would significantly limit our method and render it useless for data "in the wild".

Thus, we introduce a technique for estimating these measures based only on information about the person's pose. The prerequisite is knowledge of the actual (3D) lengths of the principal bones determined by the pose estimation model. We obtain these distances from SMPL [17]. During inference, we measure the distances in the pixel space and normalize (divide) them by their distance in the 3D space. Each measurement serves as an estimate of one SMPL model unit length in pixels, assuming that the bone is parallel to the projection plane.

We then apply simple trigonometry-based reasoning to choose the most credible estimate. Given a straight unit-length stick parallel to the ground plane, its apparent length is maximal when it is parallel to the projection plane, too, and decreases when rotating the stick around the vertical axis (down to zero when both ends visually merge to the same point). In our domain, sticks are the principal bones of different lengths. Normalizing the distances by the respective lengths makes the estimates proportional only to the cosine of the angle with the projective plane. Since cosine is a decreasing function of angle (for $\alpha \in [0°, 90°]$), the bone having the smallest angle (ideally zero) with the projection plane will correspond to the highest value. Therefore, the best estimate is the *maximum*.

Arguably, this estimate cannot be considered perfect since we have no guarantee that the assumption of parallelism actually holds. However, we are interested in determining the size of proximal regions, which do not need to match the shape of the person exactly. In fact, a minor overestimation of the size is not a problem because we do not deal with pixels in the background anyway, and it can also help us handle people with different body mass.

Therefore, we determine the best multiplication factor by tuning it using the validation data. The results are presented in the ablation study (Sec. 5.3).

4. Data

In our experiments, we rely on the DensePose COCO dataset [8]. This dataset contains about 50 thousand annotated people on a subset of images from the COCO dataset [16]. In addition to the bounding box coordinates, instance segmentation mask, and keypoints (skeleton), the ground-truth information about every instance includes the body segmentation mask and a set of dense correspondences (over 5 million annotated points in total).

The dataset is divided into train and validation splits with a ratio of about 95/5.

4.1. Assessing the quality of annotations

During our research, we repeatedly encountered incorrectly annotated instances in DensePose COCO. Therefore, as part of our efforts, we conducted research on their

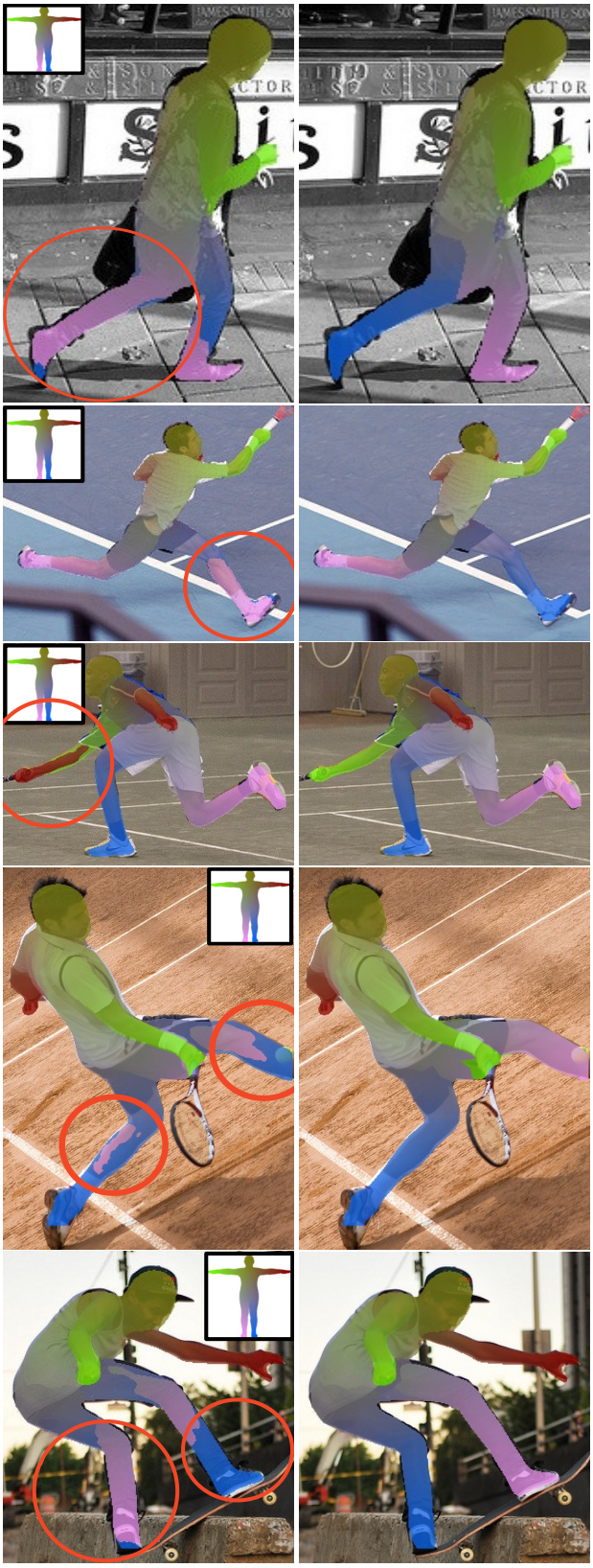


Figure 3. CSE [20] (left) vs. PC-CSE conditioned by estimated pose (right). Pose constraints ensure smoother UV maps and prevent limb duplication within a single image. A frontal view of the SMPL model [17] is shown to help assess the UV estimation.

overall quality. We define miscellaneous metrics that express the consistency of an instance's ground truth data. (For more details, see supplementary.) Then, we manually inspect the lowest-ranking instances and identify the most common problems:

1. Annotators of dense correspondences confuse the left and right parts of the body. In most cases, only one pair of body parts is confused, while the rest are annotated correctly.
2. Dense correspondences of thighs and shins are even more confused. Some instances are annotated as having only the left or only the right leg, or annotations of one leg have mixed laterality.
3. Keypoint annotators more often confuse left and right per limb or the orientation of the entire body rather than a single pair of keypoints.
4. When multiple people at least partially overlap with a bounding box, the annotated instance is different from the one that matches the dimensions of the bounding box.
5. Body segmentation masks are incomplete; not all body parts are segmented.
6. Bounding boxes lack the "is crowd" label. These are supposed to annotate many people at once (*i.e.*, a crowd) and should not be associated with dense or keypoint annotations.

We do not make any corrections to the ground truth, but we remove dense annotations that we consider wrong. We assess the precision per body part, not individually per point. If a body part shows any of the above problems, we remove all associated points regardless of laterality. As a result, we remove ca. 1.5% points from the dataset, concerning ca. 7.5% instances. (For the validation subset, the numbers are somewhat higher: 2.4% points on 11.2% instances.)

5. Experiments

In the following, we evaluate PC-CSE by simulating its use in practice. We take the `R_101_FPN_DL_soft_s1x` CSE model from the detectron2 toolbox [26] and consider it to be the baseline method. We run inference on images from the validation subset of the DensePose COCO dataset (Sec. 4) and obtain the baseline bounding boxes, instance segmentation, and pixel embeddings.

Then, we use the bounding boxes as input for top-down HPE models, which we obtain from the mmpose toolbox [5]. We choose several HPE models that differ in performance and provide different representations of human pose (see Sec. 3.1). Finally, we combine all outputs and apply our PC-CSE method and compare the accuracy of the newly produced UV maps to that of the baseline ones.

5.1. Evaluation metrics

We follow the modified COCO challenge protocol [16] that evaluates the match between predictions and ground-truth instances using Geodesic Point Similarity (GPS) and computes the algorithm's Average Precision (AP) by thresholding the GPS score [8]. We report the Average Precision for both the original dataset and the dataset without incorrect annotations (see Sec. 4.1).

HPE method	HPE	UV map	UV map†
None	—	66.2	68.8
ViTPose-b [27]	75.8	66.8	69.3
ViTPose-h [27]	79.1	67.0	69.6
ViTPose-h wb	78.6	67.3	69.8
RTMPose-l [18]	75.8	67.0	69.5
RTMPose-l wb [18]	69.5	66.7	69.3

Table 1. **AP results on the COCO dataset**. Constraining UV map estimation with 2D pose improves performance. More accurate poses lead to better UV maps. Using the whole-body (wb) skeleton further enhances performance due to better hand and foot constraints. Note that 2D Human Pose Estimation (HPE) is evaluated on a different COCO subset than UV map evaluation. Results marked with (†) are evaluated on data with ignored incorrect annotations, as detailed in Sec. 4.1.

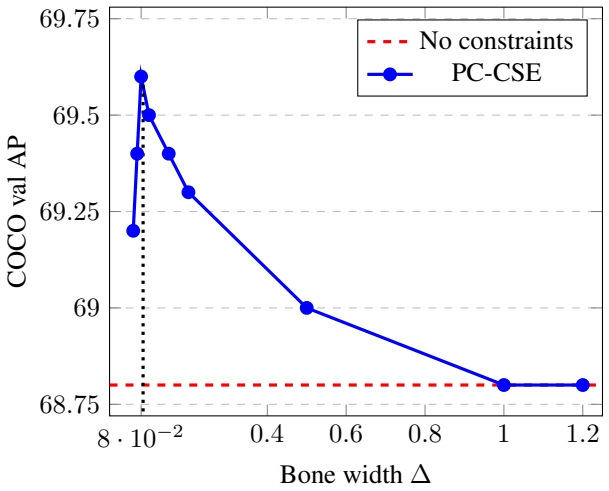


Figure 4. **Ablation on bone width** Δ defined in Sec. 3.2. RTMPose-l wb [18] is used for pose constraints. Too thin bones restrict UV Map too much and hinder performance on border pixels. Excessively thick bone estimates do not restrict UV Map sufficiently and reduce the performance gain. Note that performance with proximal regions with large regions Δ converges to the baseline method. In the extreme case when all bones are as big as the whole picture, no constraints are applied. The best value is 0.08.

5.2. Results

Table 1 compares pose-constrained CSE (PC-CSE) with the original CSE [20]. The results are reported in the COCO val dataset for comparability with previous work. Furthermore, we evaluated performance on the COCO val data set while ignoring incorrect annotations, as described in Sec. 4.1.

The first row of Tab. 1 shows the performance of CSE [20] without pose constraints. We reproduced these re-

sults and observed a 2.6 AP improvement when ignoring incorrect annotations. This gain remains consistent across all experiments.

Subsequent rows show results with pose constraints from ViTPose [27] and RTMPose [18], using different model variants. Regardless of the HPE model, applying pose-conditioned constraints consistently improves performance. As expected, the performance gain depends on the quality of the HPE model. ViTPose-h (huge) outperforms ViTPose-b (base) in HPE and achieves slightly better UV map accuracy. However, the difference is minor. Note that HPE is evaluated on a larger subset of COCO images than UV maps.

To assess the impact of the whole-body (wb) skeleton, we trained *ViTPose-h wb* on the COCO-WholeBody dataset [12]. It achieves 67.3 AP on COCO-WholeBody and 78.6 AP on COCO, compared to 79.1 AP for ViTPose-h. While the whole-body poses are less accurate, the inclusion of fingers and toes compensates for this in specific body regions.

Results for RTMPose [18] follow a similar trend. Using estimated poses improves the performance of the UV map between models, although exact gains differ. For instance, RTMPose-l matches ViTPose-b in HPE performance, but achieves slightly higher UV map accuracy. However, this difference is negligible.

RTMPose-l wb shows a much weaker HPE performance but comparable UV map accuracy. Although the inclusion of fingers and toes benefits the hand and foot regions, the reduced accuracy of other keypoints diminishes overall gains, making the trade-off less favorable.

While conditioning UV map predictions on pose significantly improves consistency, this translates to only a modest 1 AP point increase in overall performance due to several factors. The most significant issue is segmentation errors — pixels outside the segmentation mask are not assigned UV map estimates, leading to penalties. An example is shown in image Fig. 6. Detection errors also impact performance; if a person is not detected, no UV estimation can be performed.

Achieving 100 AP is challenging due to the limitations of ground truth annotations, which are human estimates often obscured by clothing. In images with loose clothing, these annotations can be highly imprecise, making it difficult to determine whether discrepancies stem from ground truth errors or model predictions. As a result, images with GPS around 80 already represent strong estimates, as shown in Fig. 6.

Examples of significant improvements over the baseline are shown in Fig. 1 and Fig. 3. These include artifact removal, better continuity between limbs, and elimination of redundant body part assignments in baseline UV maps.

5.3. Ablation study

The efficiency of PC-CSE depends on a proper outline of the proximal regions, as described in Sec. 3.2. To ensure overall robustness, we determine the best value of the bone width hyperparameter Δ by validation. We use the RTMPose-l wb model [18] and repeat the same experiment while varying the value of the hyperparameter. Note that we use the clean validation dataset that does not contain the incorrect annotations identified (Sec. 4.1).

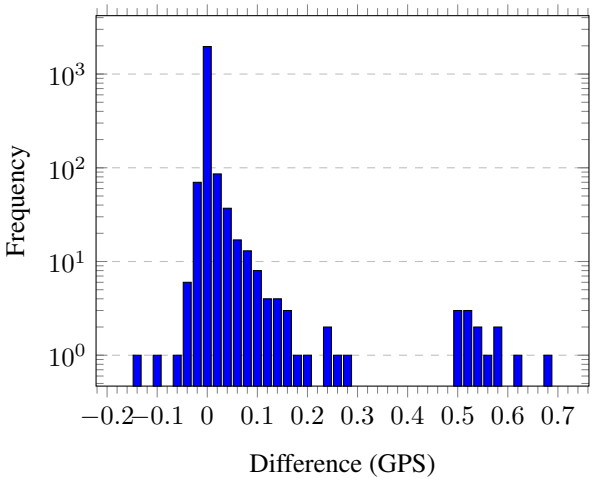


Figure 5. Image-wise performance difference of baseline CSE [20] and PC-CSE with poses from ViTPose-h wb. Performance is measured by GPS, frequency is in log scale. Positive means better performance of PC-CSE. PC-CSE causes only a few performance drops while improving more other cases, some of them dramatically, and keeping the rest about the same.

The results, shown in Fig. 4, confirm our expectations. With an increasing value of the hyperparameter, the precision increases and reaches the maximum when it is equal to 0.08. Increasing it further, we observe a gradual decrease in precision down to the baseline. This supports our earlier statement (Sec. 3.2) about the best value being a compromise and the consequences arising from a suboptimal choice. Extremely small and large values do not give our method the opportunity to have the desired impact.

Note that our experiments generally assume that the method for estimating a person's measures (Sec. 3.2) from their pose is accurate. We do not conduct any quantitative experiments on this matter, but we attempt to verify it using qualitative analysis (see supplementary material).

In addition, we provide a detailed analysis of the variation in performance metrics for each evaluated sample. An example histogram, generated for ViTPose-h wb, is shown in Fig. 5. We notice that the model maintains baseline precision on the vast majority of data samples and observe only a few performance drops, which are mainly caused by failure in the underlying pose inference. The worst are depicted in Fig. 6. However, these failures are largely compensated for by more common, sometimes drastic, improvements (shown in Fig. 3).

6. Conclusions

We presented Pose-Constrained CSE (PC-CSE), a method that conditions UV map estimation using human pose. PC-CSE leverages the robustness of 2D human pose es-

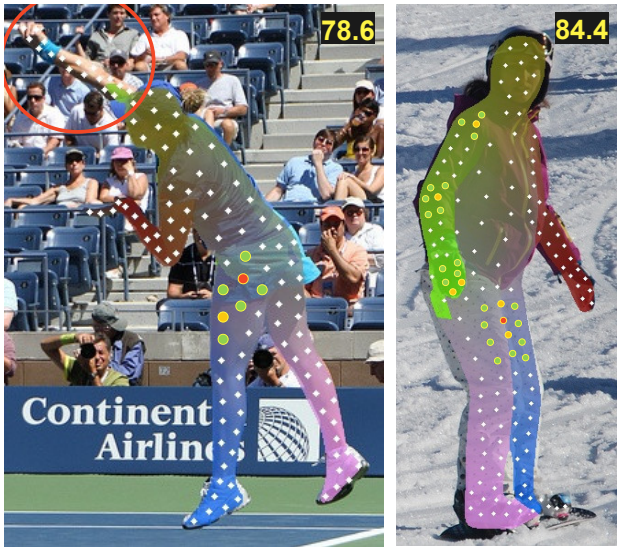


Figure 6. Images with GPS (geodesic point similarity) around 0.8. Evaluation points are shown in white. Selected wrongly estimated points (similarity < 0.5), slightly wrong (similarity $0.5 - 0.9$), and correct (similarity > 0.9). Typical errors are isolated wrong points among correct ones (left, hip), segmentation errors (left, red circle), and border points (right, legs). Loose clothing complicates annotation and estimation (right).

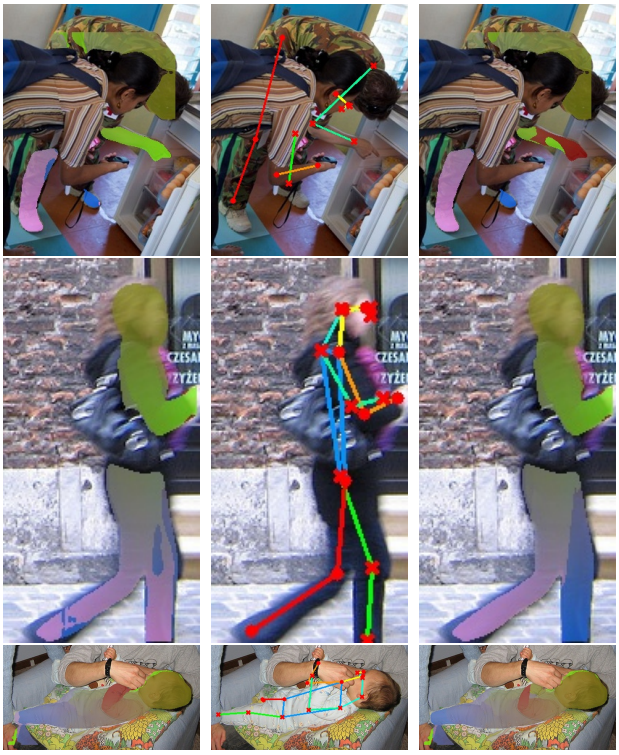


Figure 7. Three images with the largest performance decrease – CSE (left), pose estimate (middle), PC-CSE (right). Pose conditioning reduces performance when the pose estimation fails. Despite the drop, the third most negatively affected image (bottom) shows only a 0.5% decrease, highlighting that pose conditioning negatively impacts only a few images while improving many others.

timation to provide global constraints, improving the consistency of UV map predictions produced by CSE [20].

The original CSE [20] assigns pixels to vertices independently, which can lead to errors, such as assigning the same body part to multiple locations in the image and discontinuities in the same body part, as shown in Fig. 3. PC-CSE introduces global supervision through pose constraints, ensuring that while pixel assignments remain independent, the global pose structure improves the consistency of the UV map. This results in more coherent UV maps, free from artifacts and duplicated limbs.

Key findings are:

1. Conditioning UV maps by pose, even with rudimentary constraints, provides consistent improvements, though overall performance gains remain modest.
2. The choice of pose estimation model architecture has a negligible impact on the results.
3. Whole-body skeletons enable more precise constraints for hands and feet, yielding small improvements over body-only skeletons without additional computational costs.
4. COCO DensePose annotations are not entirely reliable; at least 1.5% of the points are inconsistent with pose keypoints or are otherwise inaccurate. The accuracy of points under loose clothing remains uncertain as we could neither confirm nor disprove their precision.

Limitations. The primary limitation of PC-CSE lies in its reliance on precise pose estimation. The method assumes that 2D human pose estimation (HPE) models are robust to challenges such as extreme poses, occlusions, and image deformations, which can condition UV map estimation effectively. However, if the estimated pose is inaccurate, the constrained UV map will also be incorrect. The most common errors occur in multi-body scenarios.

Another limitation arises when two body parts are in close proximity. For instance, when a person is sitting with crossed legs, pose constraints for both legs might overlap, preventing PC-CSE from correcting the original CSE estimates. Although PC-CSE does not resolve such issues, it does not degrade overall performance.

Future work. The constraints implemented by us are very coarse, as they are satisfied by letting the pixel map *somewhere* on the given body part. The corrections could become even more precise by taking the distance from its endpoints (keypoints) or the orientation of the body (frontal/dorsal) into account. In addition, there is substantial redundancy in the HPE and CSE representations, while the HPE algorithms are more advanced. The CSE method could be redesigned by building it on top of HPE and changing its objective to provide UV map estimation *given a pose estimate* (and not just the image). We also plan to use the method for UV maps on animals using SMAL [30].

Acknowledgements. This work was supported by the Ministry of the Interior of the Czech Republic project No. VJ02010041 and Czech Technical University student grant SGS23/173/OHK3/3T/13.

References

[1] SMPL - Meshcapade Wiki — meshcapade.wiki. https://meshcapade.wiki/SMPL. [Accessed 2024-12-22]. 3

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus A. Magnor. Tex2shape: Detailed full human body geometry from a single image. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2293–2303, 2019. 2

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[5] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 6

[6] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14676–14686, 2021. 2

[7] Riza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2614–2623, 2016. 2

[8] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 2, 5, 6, 1

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2

[10] Anastasia Ianina, Nikolaos Sarafianos, Yuanlu Xu, Ignacio Rocco, and Tony Tung. Bodymap: Learning full-body dense correspondence map. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13276–13285, 2022. 1, 2

[11] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12748–12757, 2021. 2, 1

[12] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 7

[13] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Zhaoen Su, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *ArXiv*, abs/2408.12569, 2024. 2

[14] Eric-Tuan Lê, Antonis Kakolyris, Petros Koutras, Himmy Tam, Efstratios Skordos, George Papandreou, Riza Alp Guler, and Iasonas Kokkinos. Meshpose: Unifying densepose and 3d body mesh reconstruction. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2405–2414, 2024. 2

[15] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Haoshu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10855–10864, 2018. 2

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 2, 3, 5, 6, 1

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 2, 3, 5

[18] Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. Rtmo: Towards high-performance one-stage real-time multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1491–1500, 2024. 2, 6, 7

[19] Natalia Neverova, James Thewlis, Riza Alp Güler, Iasonas Kokkinos, and Andrea Vedaldi. Slim densepose: Thrifty learning from sparse annotations and motion cues. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10907–10915, 2019. 2

[20] Natalia Neverova, David Novotný, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. *Advances in Neural Information Processing Systems*, 33:17258–17270, 2020. 1, 2, 3, 5, 6, 7, 8

[21] George Papandreou, Tyler Lixuan Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin P. Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *European Conference on Computer Vision*, 2018. 2

[22] Arnold Schwarz, Levente Hernadi, Felix Bießmann, and Kristian Hildebrand. The influence of faulty labels in data sets on human pose estimation. *arXiv preprint arXiv:2409.03887*, 2024. 2

[23] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[24] Feitong Tan, Danhang Tang, Mingsong Dou, Kaiwen Guo, Rohit Pandey, Cem Keskin, Ruofei Du, Deqing Sun, Sofien Bouaziz, S. Fanello, Ping Tan, and Yinda Zhang. Humangps: Geodesic preserving feature for dense human correspondences. *2021 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 1820–1830, 2021. 2

[25] TikTok. Tiktok. https://www.tiktok.com. [Accessed on 2024-12-17]. 1

[26] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 6

[27] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 6, 7

[28] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul L. Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shimin Hu. Pose2seg: Detection free human instance segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 889–898, 2018. 2

[29] Mu Zhou, Lucas Stoffl, Mackenzie W. Mathis, and Alexander Mathis. Rethinking pose estimation in crowds: overcoming the detection information bottleneck and ambiguity. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14643–14653, 2023. 2

[30] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

DOI: 10.3217/978-3-99161-022-9-003

Human Pose-Constrained UV Map Estimation

Supplementary Material

A. Annotation data quality assessment

In Sec. 4.1, we conduct research on the quality of annotations from the DensePose COCO dataset [8, 16]. In order to efficiently identify the majority of erroneous annotations without having to manually examine the entire dataset, we take the following approach.

We establish several metrics to quantify the level of (in)consistency or "(im)plausibility" of each annotated example:

- Proportion of the body segmentation covered by the instance mask. By definition, the body mask should be completely covered by the instance mask. This allows revealing problems 4 and 6, as enumerated in Sec. 4.1.
- Proportion of the area of the instance or body mask and the bounding box. Ideally, the mask should cover a significant portion of the bounding box.
  This allows revealing problems 4, 5, and 6.
- Proportion of point-wise annotations within the instance or body mask. Human segmentation should ideally contain all (visible) keypoint annotations and dense correspondences, which concern the body, too. Likewise, this allows revealing problems 4, 5, and 6.
- Ratio of median points-to-bone distances.
  We group ground-truth dense correspondences by body part and compute their median distance to the respective bone defined by ground-truth keypoint annotations (bone selection is done analogously to our mesh partitioning procedure, which exploits its resemblance to the COCO skeleton; see Sec. 3.1). We add up median distances for the same body part of either laterality. Then, we repeat the same procedure with the laterality of the keypoints flipped, and obtain another score. The ultimate value of the metric is the ratio of the two sums. When this value is high ($\gg 1$), it indicates a possibly confused laterality of keypoints or dense correspondences.
  This allows revealing problems 1 and 3.
- Inference error. We run inference on all images from the dataset and compute the mean geodesic distance (error) per body part. High inference error usually indicates deficiencies in the model's performance, but, especially on training data, it might also help reveal annotation errors. We took advantage of repeated retraining and evaluation of the inference model ("human in the loop") as it could initially have been overfitted to annotation errors.
  This allows revealing problems 1 and 2.

We sort all annotations from the least consistent and manually examine them in this order until annotations with no apparent problems start to prevail. This process is carried out individually for each defined metric.

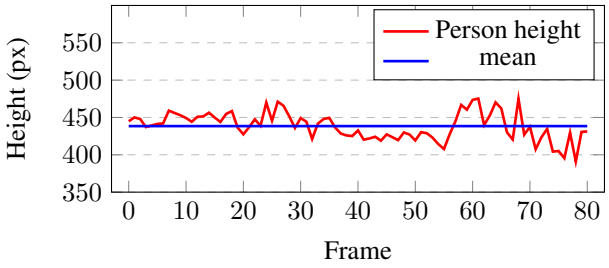


Figure 8. **Ablation on height estimation.** We infer pose from a dance video from [11] at 10 frames per second and estimate the dancing person's height in pixels (red) using the algorithm in Sec. 3.2. The variable exhibits some noise due to pose changes, but remains within the interval of a few tens of pixels at all times. The bigger noise at the end of the video is caused by more extreme poses.

B. Ablation study on height estimation

Our PC-CSE relies on estimating the proper outline of each constraint's region. In Sec. 3.2, we describe the algorithm we use to approximate the person's body measurements in pixel units of the image using only its inferred pose. The precision of such an algorithm can usually be determined by comparing the actual values and their algorithmic estimate on many images. We do not conduct such an experiment because of the lack of ground-truth data, but we verify its performance by taking a different approach.

The goal is to demonstrate that the estimate is not dramatically influenced by pose variations. However, images of people "in the wild" usually also differ in the distance of the person from the camera, as well as the underlying camera parameters. For a sensible comparison, these two factors need to remain constant. We notice that this requirement is met, for example, by short videos of people dancing in front of the camera uploaded to social networks such as TikTok [25].

Therefore, we take advantage of the TikTokDataset [11] and select several videos where a person performs a dance in front of a static camera without moving around the place. We run pose inference per video frame and record the height estimate. An example chart recording the progress of one video is shown in Fig. 8. The variable does exhibit some noise, approximately on the scale of tens of pixels, which can be attributed to pose variations and noise in the pose estimation, but it remains centered on its mean value throughout. We note that the actual noise influencing the estimate of bone width (Δ) is much smaller since the bone width is a small fraction of the person's height.