Semi-supervised Dense Keypoints using Unlabeled Multiview Images

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper presents a new end-to-end semi-supervised framework to learn a dense keypoint detector using unlabeled multiview images. A key challenge lies in finding the exact correspondences between the dense keypoints in multiple views since the inverse of keypoint mapping can be neither analytically derived nor differentiated. This limits applying existing multiview supervision approaches on sparse keypoint detection that rely on the exact correspondences. To address this challenge, we derive a new probabilistic epipolar constraint that encodes the two desired properties. (1) Soft correspondence: we define a matchability, which measures a likelihood of a point matching to the other image's corresponding point, thus relaxing the exact correspondences' requirement. (2) Geometric consistency: every point in the continuous correspondence fields must satisfy the multiview consistency collectively. We formulate a probabilistic epipolar constraint using a weighted average of epipolar errors through the matchability thereby generalizing the point-to-point geometric error to the field-to-field geometric error. This generalization facilitates learning a geometrically coherent dense keypoint detection model by utilizing a large number of unlabeled multiview images. Additionally, to prevent degenerative cases, we employ a distillation-based regularization by using a pretrained model. Finally, we design a new neural network architecture, made of twin networks, that effectively minimizes the probabilistic epipolar errors of all possible correspondences between two view images by building affinity matrices. Our method shows superior performance compared to existing methods, including non-differentiable bootstrapping in terms of keypoint accuracy, multiview consistency, and 3D reconstruction accuracy.

1 Introduction

The spatial arrangement of keypoints of dynamic organisms characterizes their complex pose, providing a computational representation of the way they behave. Recently, computer vision models offer fine grained behavioral modeling through dense keypoints that establish an injective mapping from the image coordinates to the continuous body surface of humans [8] and chimpanzees [36]. These models predict the continuous keypoint field from an image, supervised by a set of densely annotated keypoints, which shows remarkable performance on real-world imagery and brings out a number of applications including 3D mesh reconstruction [53, 52, 35, 49, 7, 21], texture/style transfer [26, 37], and geometry learning [14, 1]. Nonetheless, attaining such densely annotated data is labor intensive, and more importantly, the quality of the annotations is fundamentally bounded by the visual ambiguity of keypoints, e.g., points on textureless shirt. This visual ambiguity leads to a suboptimal model when applying it to out-of-sample distributions. In this paper, we present a new semi-supervised method to learn a dense keypoint detection model from the unlabeled multiview images.

Our main conjecture is that the dense keypoint model is optimal when it is geometrically consistent across views. That is, every pair of corresponding keypoints, independently predicted by two

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

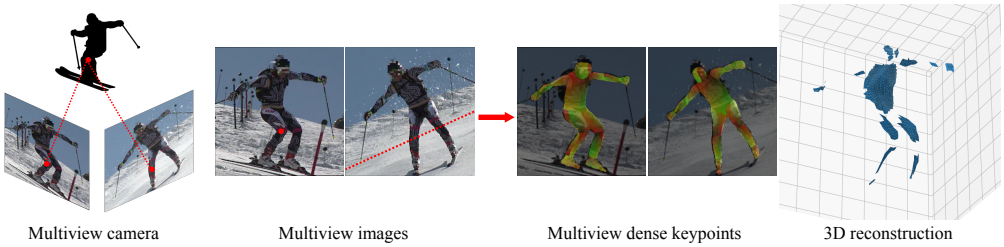

Figure 1: We use unlabeled multiview images to learn a dense keypoint model via the epipolar geometry in an end-to-end fashion. As a byproduct, we can reconstruct the 3D body surface by triangulating visible regions of body parts.

views, must satisfy the epipolar constraint [9]. However, enforcing the epipolar constraint to learn a dense keypoint model is challenging because (1) the ground truth 3D model is unknown and thus the projections of the 3D model cannot be used as the ground truth dense keypoints; (2) the predicted dense keypoints are inaccurate and continuous over the body surface, and therefore, existing multiview supervision approaches [50, 38, 45] for sparse keypoints are not applicable. In these previous methods, the epipolar constraint was enforced between two keypoints (or features) of which semantic meaning was explicitly defined by a finite set of joints (e.g., elbow channel in a network); and (3) establishing correspondences across views requires knowing an inverse mapping from the body surface to the image that can be neither analytically derived nor differentiable. These challenges limit the performance of previous work [22] that relies on iterative offline bootstrapping, which is not end-to-end trainable, or requires additional parameters to learn for 3D reconstruction[1].

We tackle these challenges through a *probabilistic epipolar constraint* by incorporating an uncertainty in correspondences. This new constraint encodes the two desired properties. (1) Soft correspondence: given a keypoint in one image, we define matchability—the likelihood of correspondence for all predicted keypoints in another image based on the geodesic distance in the body surface coordinate (e.g., texture coordinate). This allows evaluating geometric consistency in the form of a weighted average of epipolar errors over continuous body surface coordinates, eliminating the requirement of exact correspondences. (2) Geometric consistency: we generalize symmetric Sampson distance [9] for all possible pairs of keypoints from two views to enforce the epipolar constraint, collectively. With these properties, we derive a new differentiable multiview consistency measure that is label-agnostic, allowing us to utilize a large number of the unlabeled multiview images without explicit 3D reconstruction.

We design an end-to-end trainable twin network architecture that takes a pair of images as an input and outputs geometrically consistent dense keypoint fields. This network design builds the affinity maps between two keypoint fields based on the matchability and epipolar errors, which facilitates measuring the probabilistic epipolar errors for all possible correspondences. In addition, inspired by knowledge distillation, we use a pretrained model to regularize network learning, which can prevent degenerate cases. Our method shows superior performance compared to existing methods, including non-differentiable bootstrapping [22] in terms of keypoint accuracy, multiview consistency, and 3D reconstruction accuracy.

Our contributions include: (1) a novel formulation of probabilistic epipolar constraint that can be used to enforce multiview consistency on continuous dense keypoint fields in a differentiable way; (2) a new design of the neural network that enable to precisely measure the probabilistic epipolar error, which allows utilizing a large number of the unlabeled multiview images; (3) a distillation-based regularization to prevent degenerate model learning; (4) strong performance on real-world multiview image data, including Human3.6M [11], Ski-Pose [40], and OpenMonkeyPose [3], outperforming existing methods including non-differentiable dense keypoint learning [22].

Broader Impact Statement The ability to understand animals' individual and social behaviors is of central importance to multiple disciplines such as biology, neuroscience, and behavioral science. Measuring their behaviors has been extremely challenge due to limited annotated data. This approach offers a way to address this challenge with a limited number of annotated data, which will lead to a scalable behavioral analysis. The negative societal impact of this work is minimum.

[1] An analogous insight has been used for fundamental matrix, directly computed from correspondences that does not require additional variables for 3D reconstruction.

2 Related Work

Our framework aims at training a dense keypoint field estimation model via multiview supervision. We briefly review the related works.

Dense Keypoint Field Estimation Finding dense correspondence fields between two images is a challenging problem in computer vision. 3D measurements (e.g., depth and pointcloud) can provide a strong geometric cues which enable matching of deformable shapes [42, 29, 47]. Similarly, in 2D, visual and geometric cues have been used to find the dense correspondence fields in an unsupervised learning [55, 6, 4, 44, 43]. Notably, for special foreground targets such as humans, the dense matching problem can be cast as finding a dense keypoint field that maps pixel coordinates to a canonical body surface coordinates [24] (e.g., DensePose [8]). These works were built upon a large amount of data labeled by crowd-workers and generalized to learn the correspondence fields for face [2] and chimpanzees [36]. Key limitations of these approaches the inaccuracy of labeling and requirement of large labeled data. We address these limitations by formulating multiview supervision that can enforce geometric consistency, which allows utilizing a large amount of unlabeled multiview images.

Multiview Feature Learning Epipolar geometry can be used to learn a visual representation by transferring visual information from one image to another via epipolar lines or 3D reconstruction. For instance, a fusion layer can be learned to fuse feature maps across views [30]. Such fusion models can be factored into generic and camera-dependent parts to reduce the number of learnable parameters and improve generalizability [48]. Given the camera calibration, a light fusion module can be learned to directly fuse deep features [10] or heatmaps [54] from other view along corresponding epipolar line. Several works combine multiview image features to form 3D features [13, 46] or view-invariant feature in 2D [32].

Multiview Supervision Synchronized multiview images [11, 15, 51] possess a unique geometric property: images are visually similar yet geometrically distinctive, provided by stereo parallax. Such property offers a new opportunity to learn a geometrically coherent representation without labels. Bootstrapping by 3D reconstruction [22] can be used to learn a keypoint detector supervised by the projection of the 3D reconstruction to enforce cross-view consistency. MONET [50] enables an end-to-end learning by eliminating the necessity of 3D reconstruction and directly minimizing epipolar error. Learning keypoints can be combined with 3D pose estimation [34, 12] by enforcing predicting the same pose in all views while using a few labeled examples with 3D or 2D pose annotations to prevent degeneration. One can alleviate the need for large amounts of annotations by matching the predicted 3D pose with the triangulated pose [19]. Several works further learn a latent representation encoding 3D geometry from images [5, 25] or 2D pose [33, 41] by enforcing consistent embedding, texture [28], and view synthesis [45] across views. Unlike these approaches designed for sparse keypoints where the geometric consistency is applied on finite points, we study geometric consistency on continuous dense keypoint fields. Capture Dense [22] is the closest work to ours, which uses bootstrapping through 3D reconstruction of human mesh model [16]. However, due to the non-differentiable nature of bootstrapping, it is not end-to-end trainable. A temporal consistency has been also used for self-supervise the dense keypoint detector [27].

3 Method

We present a novel method to learn a dense keypoint detector by using unlabeled multiview images. We formulate the epipolar constraint for continuous correspondence fields, which allows us to enforce geometric consistency between views.

3.1 Dense Epipolar Geometry

Given a pair of synchronized images from two different views, a pair of corresponding points $\mathbf{x} \leftrightarrow \mathbf{x}'$, are related by a fundamental matrix given the calibrated cameras, i.e. $\widetilde{\mathbf{x}}'^{\mathsf{T}}\mathbf{F}\widetilde{\mathbf{x}} = 0$, where $\mathbf{F} \in \mathbb{R}^{3\times 3}$ is the fundamental matrix, and $\widetilde{\mathbf{x}} \in \mathbb{P}^2$ is a homogeneous representation of $\mathbf{x}$. The measure of geometric consistency between the two images can be written as [9]:

$$d(\mathbf{x}, \mathbf{x}'; \mathbf{F}) = \frac{|\widetilde{\mathbf{x}}'^{\mathsf{T}}\mathbf{F}\widetilde{\mathbf{x}}|}{\sqrt{(\mathbf{F}\widetilde{\mathbf{x}})_1^2 + (\mathbf{F}\widetilde{\mathbf{x}})_2^2}}, \quad \mathbf{x} \leftrightarrow \mathbf{x}' \tag{1}$$

where $(\mathbf{F}\mathbf{x})_i$ is the ith entry of $\mathbf{F}\mathbf{x}$.

Consider an injective dense keypoint mapping $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ that maps a pixel coordinate to a canonical 2D body surface coordinate, i.e., $\mathbf{u} = \phi(\mathbf{x}; \mathcal{I})$ where $\mathbf{x} \in \Theta(\mathcal{I})$ is the range of the

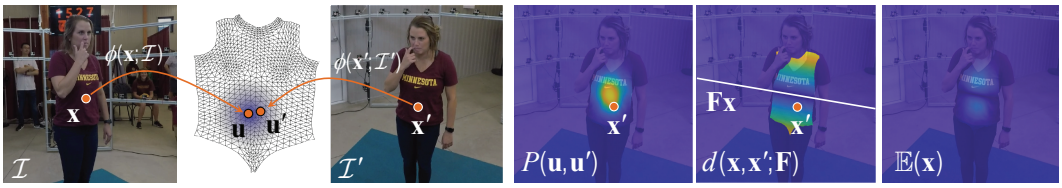

Figure 2: A dense keypoint field maps a point in an image to the canonical body surface coordinate, i.e., $\mathbf{u} = \phi(\mathbf{x};\mathcal{I})$. Establishing a correspondence between two view images requires the analytic inverse of ϕ which does not exist in general. We present a matchability $P(\mathbf{u},\mathbf{u}') = P(\phi(\mathbf{x};\mathcal{I}), \phi(\mathbf{x}';\mathcal{I}'))$, a likelihood of matching through the body surface coordinate. We combine the matchability with the epipolar constraint to obtain the probabilistic epipolar error, or expectation of geometric error $\mathbb{E}(\mathbf{x})$. It is a weighted average of epipolar errors of $\mathbf{x}$ over all possible correspondences in $\mathcal{I}'$.

foreground pixels in the image $\mathcal{I}$, and $\mathbf{u} \in \mathbb{R}^2$ is the 2D coordinate in the body surface as shown in Figure 2. One can find a correspondence between the two images through $\mathbf{u}$, i.e., $\phi^{-1}(\mathbf{u};\mathcal{I}) \leftrightarrow \phi^{-1}(\mathbf{u};\mathcal{I}')$ where $\mathcal{I}$ and $\mathcal{I}'$ the two view images. However, ϕ is an injective mapping where the analytic inverse does not exist in general.

Given a point $\mathbf{x}$ in the image $\mathcal{I}$, one can measure the expectation of geometric error by a nearest neighbor search on the body surface space:

$$\mathbb{E}(\mathbf{x}) = d(\mathbf{x},\mathbf{x}';\mathbf{F}), \quad \mathbf{x}' = \underset{\mathbf{x}'\in\Theta(\mathcal{I}')}{\operatorname{argmin}} \|\phi(\mathbf{x};\mathcal{I}) - \phi(\mathbf{x}';\mathcal{I}')\|,$$

where $\mathbb{E}(\mathbf{x})$ is the expectation of geometric error at $\mathbf{x}$. The expectation of the geometric error measures the epipolar error over all possible matches in the other view image, i.e., $\forall \mathbf{x}' \in \Theta(\mathcal{I}')$.

There are two limitations in the nearest neighbor search: (1) The correspondences are not exact, leading to a biased estimate of the geometric error expectation. For instance, a point that is not visible from another view is forced to be matched to a non-existing point, producing a large geometric error. (2) The argmin operation is not differentiable, which cannot be used to learn a dense keypoint detector in an end-to-end fashion.

Instead, we make use of a soft correspondence to address these limitations. Consider a matchability in the body surface coordinate:

$$P(\mathbf{u},\mathbf{u}') = \exp\left(-\frac{\|\mathbf{u}-\mathbf{u}'\|^2}{2\sigma^2}\right) / \sum_{\mathbf{v}\in\Omega} \exp\left(-\frac{\|\mathbf{u}-\mathbf{v}\|^2}{2\sigma^2}\right)$$

where $P(\mathbf{u},\mathbf{u}')$ is a matchability—a likelihood of $\mathbf{u}$ matched to $\mathbf{u}'$ in the body surface coordinate (Figure 2). Ω is the range of the body surface, and σ is the standard deviation that controls the smoothness of matching, e.g., when $\sigma \to 0$, it approximates the nearest neighbor search.

We use the matchability to form a probabilistic epipolar distance:

$$\mathbb{E}(\mathbf{x},\mathbf{x}') = P\left(\phi(\mathbf{x};\mathcal{I}), \phi(\mathbf{x}';\mathcal{I}')\right) d(\mathbf{x},\mathbf{x}';\mathbf{F}) \tag{2}$$

where $\mathbb{E}(\mathbf{x},\mathbf{x}')$ is the expectation of geometric error between the corresponding points.

The error expectation of $\mathbf{x}$ can be computed by marginalizing over all $\mathbf{x}'$:

$$\mathbb{E}(\mathbf{x}) = \sum_{\mathbf{x}'\in\Theta(\mathcal{I}')} \mathbb{E}(\mathbf{x},\mathbf{x}') \tag{3}$$

That is, the expectation of the geometric error of $\mathbf{x}$ measures a weighted average of epipolar errors for all possible correspondences. The higher matchability, the more contribution to the error expectation.

Given the probabilistic epipolar error over the dense keypoint field, we derive a symmetric multiview consistency loss $\mathcal{L}_{\mathrm{M}}$ as:

$$\mathcal{L}_{\mathrm{M}}(\mathcal{I},\mathcal{I}') = \frac{1}{V}\sum_{\mathbf{x}\in\Theta(\mathcal{I})} v(\mathbf{x},\mathcal{I}')\mathbb{E}(\mathbf{x}) + \frac{1}{V'}\sum_{\mathbf{x}'\in\Theta(\mathcal{I}')} v(\mathbf{x}',\mathcal{I})\mathbb{E}(\mathbf{x}') \tag{4}$$

where $v(\mathbf{x},\mathcal{I}') \in \{0,1\}$ is a visibility indicator that output one if a point $\mathbf{x}$ is visible in the other view image $\mathcal{I}'$ and zero otherwise, $V = \sum_{\mathbf{x}\in\Theta(\mathcal{I})} v(\mathbf{x},\mathcal{I}')$ and $V' = \sum_{\mathbf{x}\in\Theta(\mathcal{I})} v(\mathbf{x},\mathcal{I}')$ are the

Figure 3: Our multiview supervision progressively minimizes the epipolar error between two views (top and bottom) as learning the dense keypoint detection model. The keypoint detection, independently by a pretrained model (Iter 0), is not geometrically consistent. As the optimization progresses, the error is significantly reduced, resulting in a geometrically coherent model.

numbers of foreground pixels visible in $\mathcal{I}'$ and $\mathcal{I}$, respectively. In fact, this loss over field-to-field correspondence is a generalization of Sampson distance [9] over point-to-point correspondence.

Figure 2 illustrates the probabilistic epipolar error through the matchability. Given the dense keypoint fields from two views, $\phi(\mathbf{x};\mathcal{I}) \leftrightarrow \phi(\mathbf{x}';\mathcal{I}')$, we compute the matchability $P(\phi(\mathbf{x};\mathcal{I}), \phi(\mathbf{x}';\mathcal{I}'))$ by measuring the geodesic distance in the body surface coordinate. We combine the matchability with the epipolar constraint $d(\mathbf{x}, \mathbf{x}'; \mathbf{F})$ over all possible correspondences given $\mathbf{x}$. This results in the probabilistic epipolar error, or the expectation of geometric error for $\mathbf{x}$.

Existing approaches such as bootstrapping [22] establish the matching through 3D reconstruction of mesh and enforce the geometric error in an alternating fashion due to the non-differentiability of matching. Our differentiable formulation allows learning the dense keypoint detector in an end-to-end manner, which is flexible and shows superior performance.

3.2 Multiview Semi-supervised Learning

We learn the dense keypoint detector ϕ by minimizing the following error:

$$\mathcal{L} = \sum_{\{\mathcal{I},\mathbf{U}\}\in\mathcal{D}_{\mathrm{L}}} \mathcal{L}_{\mathrm{L}}(\mathbf{U},\mathcal{I}) + \lambda_{\mathrm{M}} \sum_{\{\mathcal{I},\mathcal{I}'\}\in\mathcal{D}_{\mathrm{P}}} \mathcal{L}_{\mathrm{M}}(\mathcal{I},\mathcal{I}') + \lambda_{\mathrm{R}} \sum_{\mathcal{I}\in\mathcal{D}_{\mathrm{U}}} \mathcal{L}_{\mathrm{R}}(\mathcal{I}) + \lambda_{\mathrm{T}} \sum_{\{\mathcal{I},\mathcal{I}'\}\in\mathcal{D}_{\mathrm{P}}} \mathcal{L}_{\mathrm{T}}(\mathcal{I},\mathcal{I}')$$

where $\mathcal{L}_{\mathrm{L}}$ is the labeled data loss, $\mathcal{L}_{\mathrm{M}}$ is the multiview supervision loss, $\mathcal{L}_{\mathrm{R}}$ is the regularization loss, and $\mathcal{L}_{\mathrm{T}}$ is the multiview photometric consistency loss. λ_{M}, λ_{R} and λ_{T} are the relative weights to control the importance of multiview consistency loss, regularization loss, and multiview photometric consistency loss respectively. $\mathcal{D}_{\mathrm{L}} = \{\mathcal{I}, \mathbf{U}\}$ is the labeled dataset where $\mathbf{U} \in \mathbb{R}^{2\times H\times W}$ specify the ground truth dense keypoint coordinate. $\mathcal{I} \in \mathcal{D}_{\mathrm{U}}$ is the unlabeled image dataset, and $\mathcal{D}_{\mathrm{P}}$ is the dataset of synchronized multiview image pairs, including labeled and unlabeled images.

Supervised Loss We learn the dense keypoint detector from the labeled data by minimizing the following error:

$$\mathcal{L}_{\mathrm{L}}(\mathbf{U},\mathcal{I}) = \sum_{\mathbf{x}\in\Theta(\mathcal{I})} \|\mathbf{U}_{\mathbf{x}} - \phi(\mathbf{x};\mathcal{I})\|_1 \tag{5}$$

where $\mathbf{U}_{\mathbf{x}}$ is the ground truth keypoint coordinate at $\mathbf{x}$. In practice, when DensePose [8] is used, the dense keypoint field is defined for each body part. We apply Equation (5) to each body part.

Distillation Based Regularization Loss Enforcing multiview consistency alone can lead to degenerate cases. For instance, consider a linear transformation in the body surface, e.g., $\mathbf{v} = \mathbf{T}\mathbf{u}$ where $\mathbf{T} \in \mathbb{R}^{2\times 2}$ is a non-singular affine transformation. Any ϕ that satisfies the following condition can be equivalent dense keypoint detector:

$$\phi(\mathbf{x};\mathcal{I}) \equiv \mathbf{T}\phi(\mathbf{x};\mathcal{I}). \tag{6}$$

This indicates that there exist an infinitely number of dense keypoint detectors that satisfy the epipolar geometry. To alleviate this geometric ambiguity, we use a distillation-based regularization using a

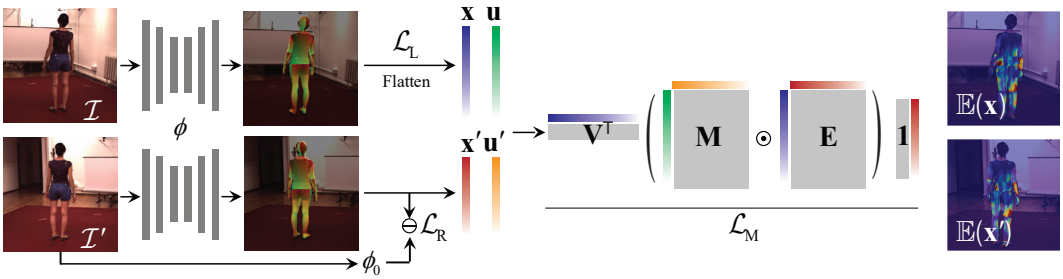

Figure 4: We design a new architecture composed of twin networks that detect dense keypoint fields. The dense keypoint fields from two views are combined to form two affinity matrices: matchability $\mathbf{M}$ and epipolar error $\mathbf{E}$. $\mathbf{M}$ is obtained from the dense keypoint fields ($\mathbf{u}$ and $\mathbf{u}'$), and $\mathbf{E}$ is obtained from the epipolar error of pixel coordinates ($\mathbf{x}$ and $\mathbf{x}'$). These matrices allow us to compute epipolar errors for all possible correspondences using probabilistic epipolar error ($\mathcal{L}_{\mathrm{M}}$). In addition, we make use of distillation-based regularization using a pretrained model ϕ_0 to avoid degenerate cases ($\mathcal{L}_{\mathrm{R}}$). We measure the labeled loss $\mathcal{L}_{\mathrm{L}}$ if the ground truth dense keypoint field is available.

pretrained model. Let ϕ_0 be a dense keypoint detector pretrained by the labeled data. We prevent the learned detector ϕ deviating too much from the pretrained detector ϕ_0 for the unlabeled data:

$$\mathcal{L}_{\mathrm{R}}(\mathcal{I}) = \sum_{\mathbf{x}\in\Theta(\mathcal{I})} \|\phi_0(\mathbf{x};\mathcal{I}) - \phi(\mathbf{x};\mathcal{I})\|^2, \tag{7}$$

where $\mathcal{L}_{\mathrm{R}}$ is the loss for the distillation-based regularization, minimizing the difference from the pretrained model. Similar to the labeled data loss, in practice, we augment the loss for body parts.

Multiview Photometric Consistency Loss We leverage photometric consistency across views. Assuming ambient light, pixels across views corresponding to the same 3D point in space should have the same RGB value. Similar to $\mathcal{L}_{\mathrm{M}}$, we minimize the photometric error as follow:

$$\mathcal{L}_{\mathrm{T}}(\mathcal{I},\mathcal{I}') = \frac{1}{V}\sum_{\mathbf{x}\in\Theta(\mathcal{I})} v(\mathbf{x},\mathcal{I}')\mathbb{T}(\mathbf{x}) + \frac{1}{V'}\sum_{\mathbf{x}'\in\Theta(\mathcal{I}')} v(\mathbf{x}',\mathcal{I})\mathbb{T}(\mathbf{x}') \tag{8}$$

where $\mathbb{T}(\mathbf{x})$ is the expectation of photometric error of $\mathbf{x}$ similar to $\mathbb{E}(\mathbf{x})$, i.e. $\mathbb{T}(\mathbf{x}) = \sum_{\mathbf{x}'\in\Theta(\mathcal{I}')} \mathbb{T}(\mathbf{x},\mathbf{x}')$ where $\mathbb{T}(\mathbf{x},\mathbf{x}') = P\left(\phi(\mathbf{x};\mathcal{I}), \phi(\mathbf{x}';\mathcal{I}')\right) \|\mathcal{I}(\mathbf{x}) - \mathcal{I}'(\mathbf{x}')\|^2$.

3.3 Network Design

We design a new network architecture composed of twin networks to learn the dense keypoint detector by enforcing multiview consistency over dense keypoint fields as shown in Figure 4. Each network is made of a fully convolutional network that outputs the dense keypoint field per body part.

Given two dense keypoint fields, we compute the probabilistic epipolar error by constructing two affinity matrices: matchability matrix and epipolar matrix, $\mathbf{M}, \mathbf{E} \in \mathbb{R}^{|\Theta(\mathcal{I})|\times|\Theta(\mathcal{I}')|}$ where $|\Theta(\mathcal{I})|$ is the cardinality of the range of foreground pixels. These two matrices are defined by:

$$\mathbf{M}_{ij} = P(\phi(\mathbf{x}_i;\mathcal{I}), \phi(\mathbf{x}_j;\mathcal{I}')), \mathbf{E}_{ij} = d(\mathbf{x}_i, \mathbf{x}_j; \mathbf{F})$$

where $\mathbf{M}_{ij}$ is the i, j entry of the matrix $\mathbf{M}$. $\mathbf{x}_i$ and $\mathbf{x}_j$ are the ith and j^{th} points from the images $\mathcal{I}$ and $\mathcal{I}'$, respectively. We also compute the visibility maps $\mathbf{V}_i = v(\mathbf{x},\mathcal{I}')$ and $\mathbf{V}'_i = v(\mathbf{x}',\mathcal{I})$.

We design a new operation to measure the multiview supervision loss in Equation (4):

$$\mathcal{L}_{\mathrm{M}} = \mathbf{V}^{\mathsf{T}}(\mathbf{M}\odot\mathbf{E})\mathbf{1}_{|\Theta(\mathcal{I}')|} + \mathbf{V}'^{\mathsf{T}}(\mathbf{M}\odot\mathbf{E})^{\mathsf{T}}\mathbf{1}_{|\Theta(\mathcal{I})|},$$

where $\mathbf{1}_n$ is the n-dimensional vector of which entries are all one.

4 Experiments

We perform experiments on human and monkey targets as two example applications to evaluate the effectiveness of our proposed semi-supervised learning pipeline.

4.1 Implementation Details

We use HRNet [18] as the backbone network followed by four head networks made up of convolutional layers to predict foreground mask, body part index, and UV coordinates on the canonical body surface, respectively. Each network takes as an input a 224×224 image and outputs 15-channel (for foreground mask head only) or 25-channel 56×56 feature maps [8].

We train the network in two stages. In the first stage, we train an initial model using the labeled data by the labeled data loss $\mathcal{L}_{\mathrm{L}}$ or using a pretrained model as the teacher network to distillate the knowledge into our network. Specifically, for application on human we follow the first approach and use 48K human instances in DensePose-COCO [8] training set, while for application on monkey we adopted the second one since we do not have access to the labelled data but only a pretrained model on chimpanzee [36]. This initial model is used for two purposes: (1) distillation-based regularization and (2) refinement via multiview supervision. For distillation-based regularization, the weights of the pretrained network is fixed. Then we leverage full loss to learn the refined network.

4.2 Evaluation Datasets

Human3.6M [11] is a large-scale indoor multiview dataset captured by 4 cameras for 3D human pose estimation. It contains 3.6 millions of images captured from 7 subjects performing 15 different daily activities, e.g. Walking, Greeting, and Discussion. Following common protocols, we use subject S1, S5, S6, S7 and S8 for training, and reserve subject S9 and S11 for testing. Following [53] and [52], we leverage SMPL parameters generated by HMR[17] via applying MoSh [23] to the sparse 3D MoCap marker data to recover ground truth 3D human meshes. We further perform Procrustes analysis [39] to align them with ground truth 3D poses in global coordinate system and then render ground truth IUV maps using PyTorch3D [31]. Since it is the only multiview human dataset that we have access to ground truth 3D mesh / IUV maps, we use this dataset to perform comprehensive experiment and studies.

Ski-Pose PTZ-Camera Dataset [40] is a multiview dataset capturing competitive skiers performing giant slalom runs. 6 synchronized and calibrated pant-tile-zoom-cameras (PTZ) cameras are used to track a single skier at a time. The global locations of the cameras were measured by a tachymeter theodolite. It contains 8.5K training images and 1.7K testing images. We use this dataset to evaluate generalization towards in-the-wild multiview settings. We use its standard train/test split to train and evaluation our model. We select 6 adjacent view pairs to form training samples.

OpenMonkeyPose [3] is a large landmark dataset of rhesus macaques captured by 62 synchronized multiview cameras. It consists of nearly 200K labeled images with four macaque subjects that freely move in a large cage while performing foraging tasks. Each monkey instance is annotated with 13 2D and 3D joints. We use this dataset to show our model's ability on transferring dense keypoints to monkey data. We split about 64K images for training and 12K images for testing. For training, we generate the densepose of monkey data using a pretrained model [36] as pseudo-labels. These pseudo-labels are then used for refining the pretrained model.

4.3 Baselines

In our experiments, we consider four baselines: (1) A model fully-supervised by labeled data, e.g. DensePose-COCO [8], which is also our initial model. We refer to this as Supervised in the following. (2) A model trained using multiview bootstrap strategy [38, 22], where multiview triangulation results from previous stage are used as the pseudo ground truth. (3) A learnable 3D mesh estimation method where dense keypoint estimation can be acquired by reprojecting 3D mesh to image domain, e.g. HMR [17]. (4) A 3D mesh estimation framework similar to HMR but additionally incorporating Model-fitting in the Loop, e.g. SPIN [20]. Supervised is also used as an initial model for our approach.

4.4 Metrics

We evaluate the performance of our dense keypoint model using metrics from three aspects: (1) geometric consistency, (2) accuracy of dense keypoints, and (3) accuracy of 3D reconstruction from multiple views.

Geometric Consistency We use epipolar distance (unit: pixel) averaged over views and frames as the metric for evaluating multiview geometric consistency. Ideally, two dense keypoints corresponding to

	Method	Keypoint accuracy				Geom. consistency	Recon. accruacy	
		AUC_{10}	AUC_{30}	mRCI	mGPS	Epi. error	MPVPE	mMVS
Comparison	Supervised	0.428	0.705	0.728	0.83	6.52	65.71	0.484
	Bootstrapping [22]	0.393	0.683	0.707	0.820	5.73	64.12	0.513
	HMR [17]	**0.495**	0.68	0.589	0.76	4.0	55.9	**0.809**
	SPIN [20]	0.456	0.615	0.44	0.685	3.26	**54.22**	0.712
	Ours	0.468	**0.72**	**0.738**	**0.834**	**2.7**	58.8	0.544
General.	Supervised (Train view 0,2 / Test view 1,3)	0.428	0.724	0.76	0.855	5.87	58.98	0.521
	Ours (Train view 0,2 / Test view 1,3)	**0.454**	**0.734**	**0.766**	**0.858**	**3.62**	**55.38**	**0.555**
	Supervised (Train view 1,3 / Test view 0,2)	0.46	0.735	0.762	0.856	5.83	57.9	0.462
	Ours (Train view 1,3 / Test view 0,2)	**0.519**	**0.756**	**0.771**	**0.861**	**3.04**	**49.94**	**0.536**
Ablation	Supervised ($\mathcal{L}_{\mathrm{L}}$)	0.48	0.748	0.777	0.864	5.54	58.71	0.521
	$\mathcal{L}_{\mathrm{L}} + \mathcal{L}_{\mathrm{M}}$	0.12	0.448	0.487	0.719	1.46	176.9	0.285
	$\mathcal{L}_{\mathrm{L}} + \mathcal{L}_{\mathrm{M}} + \mathcal{L}_{\mathrm{R}}$	0.513	0.76	0.782	0.867	2.18	55.16	0.563
	$\mathcal{L}_{\mathrm{L}} + \mathcal{L}_{\mathrm{M}} + \mathcal{L}_{\mathrm{R}} + \mathcal{L}_{\mathrm{T}}$	**0.525**	**0.764**	**0.783**	**0.868**	**2.13**	**51.17**	**0.597**

Table 1: We performance cross-method evaluation, study on model's generalizability towards new views and ablation study on Human3.6M dataset and report performance on keypoint accuracy, geometric consistency and reconstruction accuracy. (Epipolar error unit: pixel; MPVPE unit: mm)

the same point on 3D surface should have epipolar distance equal to 0. This metric can be evaluated on any multiview dataset with ground truth camera parameters available.

Dense Keypoint Accuracy We evaluate the model's performance on dense keypoint accuracy from two aspects: (1) Ratio of Correct Point (RCP) and (2) Ratio of Correct Instances (RCI). RCP evaluates correspondence accuracy over the whole image domain. Specifically, it records the ratio of foreground pixels on images with corresponding 3D body surface correctly predicted as a function of geodesic distance threshold, where the prediction is considered correct if its geodesic distance to the ground truth is below the threshold (10cm and 30cm). RCIs consider instance-wise accuracy where an instance is declared to be correct if its geodesic point similarity (GPS) [8] is above the threshold. We also report the mean RCI (mRCI) and the mean GPS for all instances (mGPS).

Reconstruction Accuracy Given dense keypoints, we measure 3D reconstruction error by triangulating them in 3D. We compute Mean Per mesh Vertex Position Error (MPVPE) as the metric for reconstruction accuracy, which is defined as the mean euclidean distance between triangulated vertices and corresponding ground truth ones. In addition, inspired by geodesic point similarity [8], we define vertex similarity as: $\mathrm{VS} = \frac{1}{|V|}\sum_{\mathbf{v}_i \in V} \exp\left(-\frac{-d(\hat{\mathbf{v}}_i - \mathbf{v}_i)^2}{2\kappa^2}\right)$, where $d(\hat{\mathbf{v}}_i - \mathbf{v}_i)$ is the euclidean distance between triangulated vertex $\hat{\mathbf{v}}_i$ and corresponding ground truth one $\mathbf{v}_i$, and V is the set of visible ground truth vertices from both views. κ is a normalizing parameter. For $\mathbf{v}_i$ that does not correspond to the triangulated vertex, it is set to infinity. Further, to account for false positives in triangulated vertices (vertices not visible from both views), we define masked vertex similarity (MVS) as $\mathrm{MVS} = \sqrt{\mathrm{VS} \cdot I}$, where I is the intersection over union between the set of triangulated vertices and V. We report mean MVS (mMVS) over all instances.

4.5 Evaluation on Human3.6M Dataset

We use Human3.6M dataset to perform (1) comprehensive cross-method evaluations, (2) study on model's generalizability towards new views, and (3) ablation study on losses used for training. Results are summarized in Table 1 with all metrics reported.

Cross-method Evaluation We evaluation the performance of our model against other methods on images from all activities of subject S9 and S11 in 10Hz. Our model and bootstrapping are trained using images from all activities of subject S1, S5, S6, S7 and S8 in 10Hz. For metrics of keypoint accuracy and geometry consistency, ours outperforms all other methods. Note that HMR and SPIN form the upper bound performance on 3D reconstruction accuracy because these models are supervised by 3D ground truth mesh model. Nevertheless, ours still achieve better performance in terms of keypoint accuracy (except for AUC_{10} only second to HMR) and geometric consistency (by a large margain of 32.5% and 17.2%).

Generalizability towards new views We evaluate generalization by testing on different views: two views are used for training and other two views are used for testing. The results are summarized in Table 1. As can be seen, although the model trained purely on one camera pair does not seen any sample captured by the other pair, its performance still get improved on all metrics on top of *baseline* model by a margin of 0.6%-12.8% on keypoint accuracy, 38.3% - 47.9% on geometric consistency, and 6.5% - 13.7% on reconstruction accuracy. This shows that model trained by our semi-supervised approach can be generalized to new views.

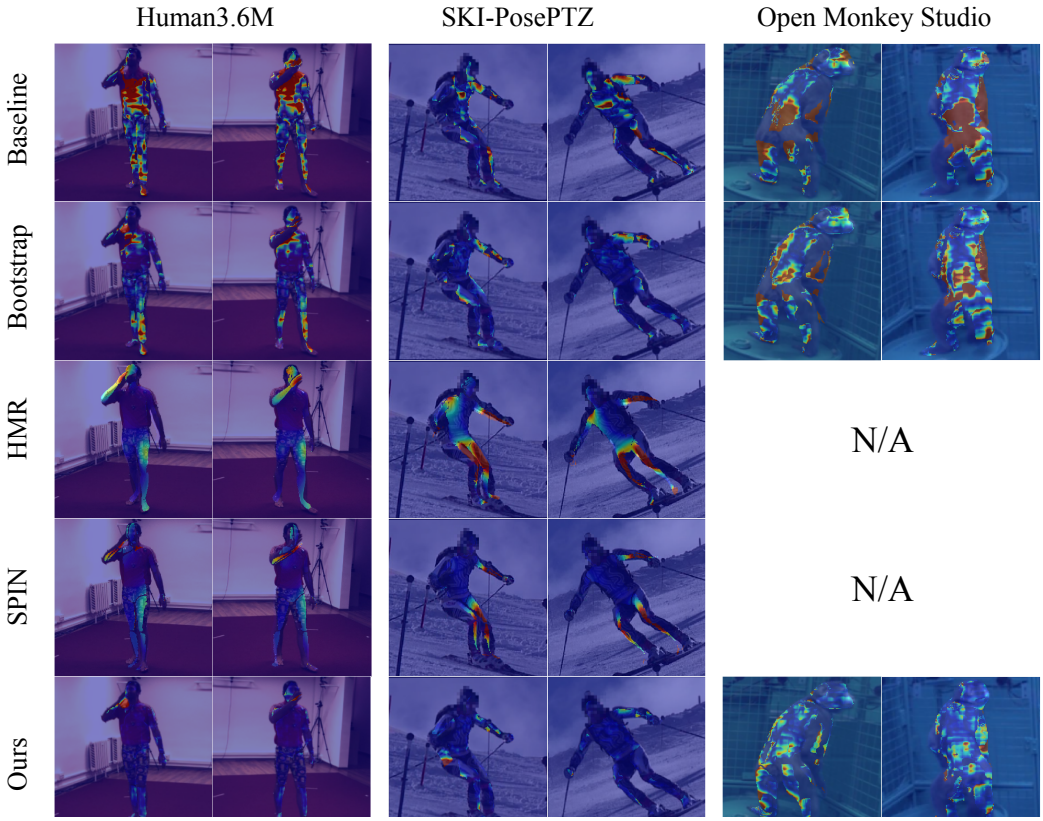

Figure 5: Qualitative results on Human3.6M, Ski-Pose PTZ-Camera and OpenMonkeyPose Datasets. Heatmaps overlapping on images indicate epipolar error for each pixels.

Ablation Study We conduct an ablation study to evaluate the impact of each loss. The results are reported in Table 1. It proves that the model trained with $\mathcal{L}_{\mathrm{L}}$ + $\mathcal{L}_{\mathrm{M}}$ shows inferior performance because of degenerate cases (Section 3.2). This limitation can be addressed by the regularization $\mathcal{L}_{\mathrm{R}}$, and further, $\mathcal{L}_{\mathrm{T}}$ improves the performance, in particular, for reconstruction accuracy.

4.6 Evalution on Ski-Pose PTZ-Camera Dataset and OpenMonkeyPose Dataset

We evaluate our method on multiview in-the-wild datasets: Ski-pose and OpenMonkeyPose. Since no ground truth is available, we evaluate geometric consistency summarized in Table 4.6 and show qualitative results (2nd and 3rd columns in Figure 5). The results show that our model outperforms other baselines by a large margin: 37.2%-54.1% on Ski-Pose and 31.9%-47.5% on OpenMonkeyPose, which can be visually identified by qualitative results.

Method	Ski-Pose [40]	Monkey [3]
Supervised	11.38	11.83
Bootstrap [22]	8.64	9.12
HMR [17]	13.4	N/A
SPIN [20]	11.83	N/A
Our	**5.43**	**6.21**

Table 2: Comparison on geometry consistency for in-the-wild data (Epipolar error unit: pixel).

5 Conclusion

We present a novel end-to-end semi-supervised approach to learn a dense keypoint detector by leverage a large amount of unlabeled multiview images. Due to the nature of continuous keypoint representation, finding exact correspondences between views is challenging unlike sparse keypoints. We address this challenge by formulating a new probabilistic epipolar constraint that allows measuring a field-to-field geometric error without knowing exact correspondences. Additionally, we proposed a distillation-based regularization to prevent degenerated cases. We design a new network architecture made of twin networks that can effectively measure the probabilistic epipolar error by considering all possible correspondences using affinity matrices. We show that our method outperforms the baseline approaches in keypoint accuracy, multiview consistency, and reconstruction accuracy.

References

[1] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019.

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

[3] P. C. Bala, B. R. Eisenreich, S. B. M. Yoo, B. Y. Hayden, H. S. Park, and J. Zimmermann. Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature Communications*, 2020.

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

[5] X. Chen, K.-Y. Lin, W. Liu, C. Qian, and L. Lin. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *CVPR*, 2019.

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

[7] R. A. Guler and I. Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, 2019.

[8] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018.

[9] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[10] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu. Epipolar transformers. In *CVPR*, 2020.

[11] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2013.

[12] U. Iqbal, P. Molchanov, and J. Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *CVPR*, 2020.

[13] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. In *ICCV*, 2019.

[14] Y. Jafarian and H. S. Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *CVPR*, 2021.

[15] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 2017.

[16] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018.

[17] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

[18] D. M. B. X. D. L. Z. Z. J. W. Ke Sun, Zigang Geng. Bottom-up human pose estimation by ranking heatmap-guided adaptive keypoint estimates. *arXiv*, 2020.

[19] M. Kocabas, S. Karagoz, and E. Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *CVPR*, 2019.

[20] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[21] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019.

[22] X. Li, Y. Liu, H. Joo, Q. Dai, and Y. Sheikh. Capture dense: Markerless motion capture meets dense pose estimation. *arXiv*, 2018.

[23] M. Loper, N. Mahmood, and M. J. Black. Mosh: Motion and shape capture from sparse markers. *TOG*, 2014.

[24] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *TOG*, 2015.

[25] R. Mitra, N. B. Gundavarapu, A. Sharma, and A. Jain. Multiview-consistent semi-supervised learning for 3d human pose estimation. In *CVPR*, 2020.

[26] N. Neverova, R. A. Guler, and I. Kokkinos. Dense pose transfer. In *ECCV*, 2018.

[27] N. Neverova, J. Thewlis, R. A. Guler, I. Kokkinos, and A. Vedaldi. Slim densepose: Thrifty learning from sparse annotations and motion cues. In *CVPR*, 2019.

[28] G. Pavlakos, N. Kolotouros, and K. Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019.

[29] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon. Metric regression forests for correspondence estimation. *IJCV*, 2015.

[30] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng. Cross view fusion for 3d human pose estimation. In *ICCV*, 2019.

[31] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv*, 2020.

[32] E. Remelli, S. Han, S. Honari, P. Fua, and R. Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *CVPR*, 2020.

[33] H. Rhodin, M. Salzmann, and P. Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *ECCV*, 2018.

[34] H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, and P. Fua. Learning monocular 3d human pose estimation from multi-view images. In *CVPR*, 2018.

[35] Y. Rong, Z. Liu, C. Li, K. Cao, and C. C. Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *ICCV*, 2019.

[36] A. Sanakoyeu, V. Khalidov, M. S. McCarthy, A. Vedaldi, and N. Neverova. Transferring dense pose to proximal animal classes. In *CVPR*, 2020.

[37] A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov, et al. Textured neural avatars. In *CVPR*, 2019.

[38] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[39] O. Sorkine-Hornung and M. Rabinovich. Least-squares rigid motion using svd. *Computing*, 2017.

[40] J. Spörri. Reasearch dedicated to sports injury prevention-the'sequence of prevention'on the example of alpine ski racing. *Habilitation with Venia Docendi in Biomechanics*, 2016.

[41] J. J. Sun, J. Zhao, L.-C. Chen, F. Schroff, H. Adam, and T. Liu. View-invariant probabilistic embedding for human pose. In *ECCV*, 2020.

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

[43] J. Thewlis, S. Albanie, H. Bilen, and A. Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *ICCV*, 2019.

[44] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. *arXiv*, 2017.

[45] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017.

[46] H. Tu, C. Wang, and W. Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. *arXiv*, 2020.

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

[48] R. Xie, C. Wang, and Y. Wang. Metafuse: A pre-trained fusion model for human pose estimation. In *CVPR*, 2020.

[49] Y. Xu, S.-C. Zhu, and T. Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *ICCV*, 2019.

[50] Y. Yao, Y. Jafarian, and H. S. Park. Monet: Multiview semi-supervised keypoint detection via epipolar divergence. In *ICCV*, 2019.

[51] Z. Yu, J. S. Yoon, I. K. Lee, P. Venkatesh, J. Park, J. Yu, and H. S. Park. Humbi: A large multiview dataset of human body expressions. In *CVPR*, 2020.

[52] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang. 3d human mesh regression with dense correspondence. In *CVPR*, 2020.

[53] H. Zhang, J. Cao, G. Lu, W. Ouyang, and Z. Sun. Learning 3d human shape and pose from dense body parts. *TPAMI*, 2020.

[54] Z. Zhang, C. Wang, W. Qiu, W. Qin, and W. Zeng. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *IJCV*, 2020.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No] See Supplementary materials.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Section 1
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Code, data, instruction will be available at the time of publication.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section 4.1 and supplementary materials
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] See Supplementary materials.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] We use publicly open dataset.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] We use publicly open dataset.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]