View-Invariant Probabilistic Embedding for Human Pose

Jennifer J. Sun[1,*], Jiaping Zhao[2], Liang-Chieh Chen[2], Florian Schroff[2], Hartwig Adam[2], and Ting Liu[2]

[1] California Institute of Technology
jjsun@caltech.edu
[2] Google Research
{jiapingz,lcchen,fschroff,hadam,liuti}@google.com

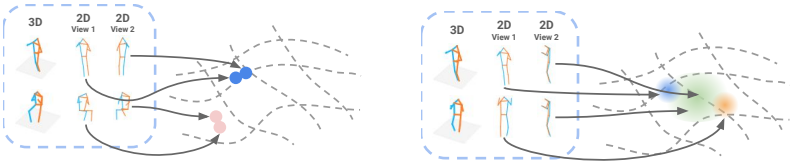
Abstract. Depictions of similar human body configurations can vary with changing viewpoints. Using only 2D information, we would like to enable vision algorithms to recognize similarity in human body poses across multiple views. This ability is useful for analyzing body movements and human behaviors in images and videos. In this paper, we propose an approach for learning a compact view-invariant embedding space from 2D joint keypoints alone, without explicitly predicting 3D poses. Since 2D poses are projected from 3D space, they have an inherent ambiguity, which is difficult to represent through a deterministic mapping. Hence, we use probabilistic embeddings to model this input uncertainty. Experimental results show that our embedding model achieves higher accuracy when retrieving similar poses across different camera views, in comparison with 2D-to-3D pose lifting models. We also demonstrate the effectiveness of applying our embeddings to view-invariant action recognition and video alignment. We plan to release our code for research.

Keywords: Human Pose Embedding, Probabilistic Embedding, View-Invariant Pose Retrieval

1 Introduction

When we represent three dimensional (3D) human bodies in two dimensions (2D), the same human pose can appear different across camera views. There can be significant visual variations from a change in viewpoint due to changing relative depth of body parts and self-occlusions. Despite these variations, humans have the ability to recognize similar 3D human body poses in images and videos. This ability is useful for computer vision tasks where changing viewpoints should not change the labels of the task. We explore how we can embed 2D visual information of human poses to be consistent across camera views. We show that these embeddings are useful for tasks such as view-invariant pose retrieval, action recognition, and video alignment.

* This work was done during the author's internship at Google.



(a) View-Invariant Pose Embeddings (VIPE).

(b) Probabilistic View-Invariant Pose Embeddings (Pr-VIPE).

Fig. 1: We embed 2D poses such that our embeddings are (a) view-invariant (2D projections of similar 3D poses are embedded close) and (b) probabilistic (embeddings are distributions that cover different 3D poses projecting to the same input 2D pose).

Inspired by 2D-to-3D lifting models [32], we learn view invariant embeddings directly from 2D pose keypoints. As illustrated in Fig. 1, we explore whether view invariance of human bodies can be achieved from 2D poses alone, without predicting 3D pose. Typically, embedding models are trained from images using deep metric learning techniques [35, 14, 8]. However, images with similar human poses can appear different because of changing viewpoints, subjects, backgrounds, clothing, etc. As a result, it can be difficult to understand errors in the embedding space from a specific factor of variation. Furthermore, multi-view image datasets for human poses are difficult to capture in the wild with 3D groundtruth annotations. In contrast, our method leverages existing 2D keypoint detectors: using 2D keypoints as inputs allows the embedding model to focus on learning view invariance. Our 2D keypoint embeddings can be trained using datasets in lab environments, while having the model generalize to in-the-wild data. Additionally, we can easily augment training data by synthesizing multi-view 2D poses from 3D poses through perspective projection.

Another aspect we address is input uncertainty. The input to our embedding model is 2D human pose, which has an inherent ambiguity. Many valid 3D poses can project to the same or very similar 2D pose [1]. This input uncertainty is difficult to represent using deterministic mappings to the embedding space (point embeddings) [37, 24]. Our embedding space consists of probabilistic embeddings based on multivariate Gaussians, as shown in Fig. 1b. We show that the learned variance from our method correlates with input 2D ambiguities. We call our approach Pr-VIPE for **Pr**obabilistic **V**iew-**I**nvariant **P**ose **E**mbeddings. The non-probabilistic, point embedding formulation will be referred to as VIPE.

We show that our embedding is applicable to subsequent vision tasks such as pose retrieval [35, 21], video alignment [11], and action recognition [60, 18]. One direct application is pose-based image retrieval. Our embedding enables users to search images by fine-grained pose, such as jumping with hands up, riding bike with one hand waving, and many other actions that are potentially difficult to pre-define. The importance of this application is further highlighted by works such as [35, 21]. Compared with using 3D keypoints with alignment for retrieval, our embedding enables efficient similarity comparisons in Euclidean space.

Contributions Our main contribution is the method for learning an embedding space where 2D pose embedding distances correspond to their similarities in absolute 3D pose space. We also develop a probabilistic formulation that captures 2D pose ambiguity. We use cross-view pose retrieval to evaluate the view-invariant property: given a monocular pose image, we retrieve the same pose from different views without using camera parameters. Our results suggest 2D poses are sufficient to achieve view invariance without image context, and we do not have to predict 3D pose coordinates to achieve this. We also demonstrate the use of our embeddings for action recognition and video alignment.

2 Related Work

Metric Learning We are working to understand similarity in human poses across views. Most works that aim to capture similarity between inputs generally apply techniques from metric learning. Objectives such as contrastive loss (based on pair matching) [4, 12, 37] and triplet loss (based on tuple ranking) [56, 50, 57, 13] are often used to push together/pull apart similar/dissimilar examples in embedding space. The number of possible training tuples increases exponentially with respect to the number of samples in the tuple, and not all combinations are equally informative. To find informative training tuples, various mining strategies are proposed [50, 58, 38, 13]. In particular, semi-hard triplet mining has been widely used [50, 58, 42]. This mining method finds negative examples that are fairly hard as to be informative but not too hard for the model. The hardness of a negative sample is based on its embedding distance to the anchor. Commonly, this distance is the Euclidean distance [56, 57, 50, 13], but any differentiable distance function could be applied [13]. [16, 19] show that alternative distance metrics also work for image and object retrieval.

In our work, we learn a mapping from Euclidean embedding distance to a probabilistic similarity score. This probabilistic similarity captures closeness in 3D pose space from 2D poses. Our work is inspired by the mapping used in soft contrastive loss [37] for learning from an occluded N-digit MNIST dataset.

Most of the papers discussed above involve deterministically mapping inputs to point embeddings. There are works that also map inputs to probabilistic embeddings. Probabilistic embeddings have been used to model specificity of word embeddings [55], uncertainty in graph representations [3], and input uncertainty due to occlusion [37]. We will apply probabilistic embeddings to address inherent ambiguities in 2D pose due to 3D-to-2D projection.

Human Pose Estimation 3D human poses in a global coordinate frame are view-invariant, since images across views are mapped to the same 3D pose. However, as mentioned by [32], it is difficult to infer the 3D pose in an arbitrary global frame since any changes to the frame does not change the input data. Many approaches work with poses in the camera coordinate system [32, 6, 43, 46, 62, 52, 48, 53, 7], where the pose description changes based on viewpoint. While our work focuses on images with a single person, there are other works focusing on describing poses of multiple people [47].

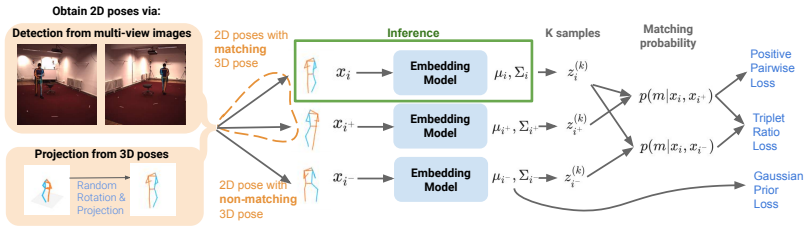


Fig. 2: Overview of Pr-VIPE model training and inference. Our model takes keypoint input from a single 2D pose (detected from images and/or projected from 3D poses) and predicts embedding distributions. Three losses are applied during training.

Our approach is similar in setup to existing 3D lifting pose estimators [32, 6, 43, 46, 9] in terms of using 2D pose keypoints as input. The difference is that lifting models are trained to regress to 3D pose keypoints, while our model is trained using metric learning and outputs an embedding distribution. Some recent works also use multi-view datasets to predict 3D poses in the global coordinate frame [44, 26, 20, 49, 54]. Our work differs from these methods with our goal (view-invariant embeddings), task (cross-view pose retrieval), and approach (metric learning). Another work on pose retrieval [35] embeds images with similar 2D poses in the same view close together. Our method focuses on learning view invariance, and we also differ from [35] in method (probabilistic embeddings).

View Invariance and Object Retrieval When we capture a 3D scene in 2D as images or videos, changing the viewpoint often does not change other properties of the scene. The ability to recognize visual similarities across viewpoints is helpful for a variety of vision tasks, such as motion analysis [23, 22], tracking [39], vehicle and human re-identification [8, 61], object classification and retrieval [27, 15, 14], and action recognition [45, 29, 59, 28].

Some of these works focus on metric learning for object retrieval. Their learned embedding spaces place different views of the same object class close together. Our work on human pose retrieval differs in a few ways. Our labels are continuous 3D poses, whereas in object recognition tasks, each embedding is associated with a discrete class label. Furthermore, we embed 2D poses, while these works embed images. Our approach allows us to investigate the impact of input 2D uncertainty with probabilistic embeddings and explore confidence measures to cross-view pose retrieval. We hope that our work provides a novel perspective on view invariance for human poses.

3 Our Approach

The training and inference framework of Pr-VIPE is illustrated in Fig. 2. Our goal is to embed 2D poses such that distances in the embedding space correspond to similarities of their corresponding absolute 3D poses in Euclidean space. We achieve this view invariance property through our triplet ratio loss (Section 3.2), which pushes together/pull apart 2D poses corresponding to similar/dissimilar 3D poses. The positive pairwise loss (Section 3.3) is applied to

increase the matching probability of similar poses. Finally, the Gaussian prior loss (Section 3.4) helps regularize embedding magnitude and variance.

3.1 Matching Definition

The 3D pose space is continuous, and two 3D poses can be trivially different without being identical. We define two 3D poses to be matching if they are visually similar regardless of viewpoint. Given two sets of 3D keypoints $(\boldsymbol{y}_i, \boldsymbol{y}_j)$, we define a matching indicator function

$$m_{ij} = \begin{cases} 1, & \text{if NP-MPJPE}(\boldsymbol{y}_i, \boldsymbol{y}_j) \leqslant \kappa \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where κ controls visual similarity between matching poses. Here, we use mean per joint position error (MPJPE) [17] between the two sets of 3D pose keypoints as a proxy to quantify their visual similarity. Before computing MPJPE, we normalize the 3D poses and apply Procrustes alignment between them. The reason is that we want our model to be view-invariant and to disregard rotation, translation, or scale differences between 3D poses. We refer to this normalized, Procrustes aligned MPJPE as **NP-MPJPE**.

3.2 Triplet Ratio Loss

The triplet ratio loss aims to embed 2D poses based on the matching indicator function (1). Let n be the dimension of the input 2D pose keypoints $\boldsymbol{x}$, and d be the dimension of the output embedding. We would like to learn a mapping $f : \mathbb{R}^n \to \mathbb{R}^d$, such that $D(\boldsymbol{z}_i, \boldsymbol{z}_j) < D(\boldsymbol{z}_i, \boldsymbol{z}_{j'}), \forall m_{ij} > m_{ij'}$, where $\boldsymbol{z} = f(\boldsymbol{x})$, and $D(\boldsymbol{z}_i, \boldsymbol{z}_j)$ is an embedding space distance measure.

For a pair of input 2D poses $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, we define $p(m|\boldsymbol{x}_i, \boldsymbol{x}_j)$ to be the probability that their corresponding 3D poses $(\boldsymbol{y}_i, \boldsymbol{y}_j)$ match, that is, they are visually similar. While it is difficult to define this probability directly, we propose to assign its values by estimating $p(m|\boldsymbol{z}_i, \boldsymbol{z}_j)$ via metric learning. We know that if two 3D poses are identical, then $p(m|\boldsymbol{x}_i, \boldsymbol{x}_j) = 1$, and if two 3D poses are sufficiently different, $p(m|\boldsymbol{x}_i, \boldsymbol{x}_j)$ should be small. For any given input triplet $(\boldsymbol{x}_i, \boldsymbol{x}_{i^+}, \boldsymbol{x}_{i^-})$ with $m_{i,i^+} > m_{i,i^-}$, we want

$$\frac{p(m|\boldsymbol{z}_i, \boldsymbol{z}_{i^+})}{p(m|\boldsymbol{z}_i, \boldsymbol{z}_{i^-})} \geqslant \beta, \tag{2}$$

where $\beta > 1$ represents the ratio of the matching probability of a similar 3D pose pair to that of a dissimilar pair. Applying negative logarithm to both sides, we have

$$(-\log p(m|\boldsymbol{z}_i, \boldsymbol{z}_{i^+})) - (-\log p(m|\boldsymbol{z}_i, \boldsymbol{z}_{i^-})) \leqslant -\log \beta. \tag{3}$$

Notice that the model can be trained to satisfy this with the triplet loss framework [50]. Given batch size N, we define triplet ratio loss $\mathcal{L}_{\text{ratio}}$ as

$$\mathcal{L}_{\text{ratio}} = \sum_{i=1}^{N} \max(0, D_m(\boldsymbol{z}_i, \boldsymbol{z}_{i^+}) - D_m(\boldsymbol{z}_i, \boldsymbol{z}_{i^-}) + \alpha)), \tag{4}$$

with distance kernel $D_m(\boldsymbol{z}_i, \boldsymbol{z}_j) = -\log p(m|\boldsymbol{z}_i, \boldsymbol{z}_j)$ and margin $\alpha = \log \beta$. To form a triplet $(\boldsymbol{x}_i, \boldsymbol{x}_{i^+}, \boldsymbol{x}_{i^-})$, we set the anchor $\boldsymbol{x}_i$ and positive $\boldsymbol{x}_{i^+}$ to be projected from the same 3D pose and perform online semi-hard negative mining [50] to find $\boldsymbol{x}_{i^-}$.

It remains for us to compute matching probability using our embeddings. To compute $p(m|\boldsymbol{z}_i, \boldsymbol{z}_j)$, we use the formulation proposed by [37]:

$$p(m|\boldsymbol{z}_i, \boldsymbol{z}_j) = \sigma(-a||\boldsymbol{z}_i - \boldsymbol{z}_j||_2 + b), \tag{5}$$

where σ is a sigmoid function, and the trainable scalar parameters $a > 0$ and $b \in \mathbb{R}$ calibrate embedding distances to probabilistic similarity.

3.3 Positive Pairwise Loss

The positive pairs in our triplets have identical 3D poses. We would like them to have high matching probabilities, which can be encouraged by adding the positive pairwise loss

$$\mathcal{L}_{\text{positive}} = \sum_{i=1}^{N} -\log p(m|\boldsymbol{z}_i, \boldsymbol{z}_{i^+}). \tag{6}$$

The combination of $\mathcal{L}_{\text{ratio}}$ and $\mathcal{L}_{\text{positive}}$ can be applied to training point embedding models, which we refer to as VIPE in this paper.

3.4 Probabilistic Embeddings

In this section, we discuss the extension of VIPE to the probabilistic formulation Pr-VIPE. The inputs to our model, 2D pose keypoints, are inherently ambiguous, and there are many valid 3D poses projecting to similar 2D poses [1]. This input uncertainty can be difficult to model using point embeddings [24, 37]. We investigate representing this uncertainty using distributions in the embedding space by mapping 2D poses to probabilistic embeddings: $\boldsymbol{x} \to p(\boldsymbol{z}|\boldsymbol{x})$. Similar to [37], we extend the input matching probability (5) to using probabilistic embeddings as $p(m|\boldsymbol{x}_i, \boldsymbol{x}_j) = \int p(m|\boldsymbol{z}_i, \boldsymbol{z}_j)p(\boldsymbol{z}_i|\boldsymbol{x}_i)p(\boldsymbol{z}_j|\boldsymbol{x}_j)\mathrm{d}\boldsymbol{z}_i\mathrm{d}\boldsymbol{z}_j$, which can be approximated using Monte-Carlo sampling with K samples drawn from each distribution as

$$p(m|\boldsymbol{x}_i, \boldsymbol{x}_j) \approx \frac{1}{K^2} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} p(m|\boldsymbol{z}_i^{(k_1)}, \boldsymbol{z}_j^{(k_2)}). \tag{7}$$

We model $p(\boldsymbol{z}|\boldsymbol{x})$ as a d-dimensional Gaussian with a diagonal covariance matrix. The model outputs mean $\mu(\boldsymbol{x}) \in \mathbb{R}^d$ and covariance $\Sigma(\boldsymbol{x}) \in \mathbb{R}^d$ with shared base network and different output layers. We use the reparameterization trick [25] during sampling.

In order to prevent variance from collapsing to zero and to regularize embedding mean magnitudes, we place a unit Gaussian prior on our embeddings with KL divergence by adding the Gaussian prior loss

$$\mathcal{L}_{\text{prior}} = \sum_{i=1}^{N} D_{\text{KL}}(\mathcal{N}(\mu(\boldsymbol{x}_i), \Sigma(\boldsymbol{x}_i)) \,\|\, \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})). \tag{8}$$

Inference At inference time, our model takes a single 2D pose (either from detection or projection) and outputs the mean and the variance of the embedding Gaussian distribution.

3.5 Camera Augmentation

Our triplets can be made of detected and/or projected 2D keypoints as shown in Fig. 2. When we train only with detected 2D keypoints, we are constrained to the camera views in training images. To reduce overfitting to these camera views, we perform camera augmentation by generating triplets using detected keypoints alongside projected 2D keypoints at random views.

To form triplets using multi-view image pairs, we use detected 2D keypoints from different views as anchor-positive pairs. To use projected 2D keypoints, we perform two random rotations to a normalized input 3D pose to generate two 2D poses from different views for anchor/positive. Camera augmentation is then performed by using a mixture of detected and projected 2D keypoints. We find that training using camera augmentation can help our models learn to generalize better to unseen views (Section 4.2.2).

3.6 Implementation Details

We normalize 3D poses similar to [7], and we perform instance normalization to 2D poses. The backbone network architecture for our model is based on [32]. We use $d = 16$ as a good trade-off between embedding size and accuracy. To weigh different losses, we use $w_{\text{ratio}} = 1$, $w_{\text{positive}} = 0.005$, and $w_{\text{prior}} = 0.001$. We choose $\beta = 2$ for the triplet ratio loss margin and $K = 20$ for the number of samples. The matching NP-MPJPE threshold is $\kappa = 0.1$ for all training and evaluation. Our approach does not rely on a particular 2D keypoint detector, and we use PersonLab [40] for our experiments. For random rotation in camera augmentation, we uniformly sample azimuth angle between $\pm 180°$, elevation between $\pm 30°$, and roll between $\pm 30°$. Our implementation is in TensorFlow, and all the models are trained with CPUs. More details and ablation studies on hyperparamters are provided in the supplementary materials.

4 Experiments

We demonstrate the performance of our model through pose retrieval across different camera views (Section 4.2). We further show our embeddings can be directly applied to downstream tasks, such as action recognition (Section 4.3.1) and video alignment (Section 4.3.2), without any additional training.

4.1 Datasets

For all the experiments in this paper, we only train on a subset of the Human3.6M [17] dataset. For pose retrieval experiments, we validate on the Human3.6M hold-out set and test on another dataset (MPI-INF-3DHP [33]), which

is unseen during training and free from parameter tuning. We also present qualitative results on MPII Human Pose [2], for which 3D groundtruth is not available. Additionally, we directly use our embeddings for action recognition and sequence alignment on Penn Action [60].

Human3.6M (H3.6M) H3.6M is a large human pose dataset recorded from 4 chest level cameras with 3D pose groundtruth. We follow the standard protocol [32]: train on Subject 1, 5, 6, 7, and 8, and hold out Subject 9 and 11 for validation. For evaluation, we remove near-duplicate 3D poses within 0.02 NP-MPJPE, resulting in a total of 10910 evaluation frames per camera. This process is camera-consistent, meaning if a frame is selected under one camera, it is selected under all cameras, so that the perfect retrieval result is possible.

MPI-INF-3DHP (3DHP) 3DHP is a more recent human pose dataset that contains 14 diverse camera views and scenarios, covering more pose variations than H3.6M [33]. We use 11 cameras from this dataset and exclude the 3 cameras with overhead views. Similar to H3.6M, we remove near-duplicate 3D poses, resulting in 6824 frames per camera. We use all 8 subjects from the train split of 3DHP. **This dataset is only used for testing.**

MPII Human Pose (2DHP) This dataset is commonly used in 2D pose estimation, containing 25K images from YouTube videos. Since groundtruth 3D poses are not available, we show qualitative results on this dataset.

Penn Action This dataset contains 2326 trimmed videos for 15 pose-based actions from different views. We follow the standard protocol [36] for our action classification and video alignment experiments.

4.2 View-Invariant Pose Retrieval

Given multi-view human pose datasets, we query using detected 2D keypoints from one camera view and find the nearest neighbors in the embedding space from a different camera view. We iterate through all camera pairs in the dataset as query and index. Results averaged across all cameras pairs are reported.

4.2.1 Evaluation Procedure We report Hit@k with $k = 1$, 10, and 20 on pose retrievals, which is the percentage of top-k retrieved poses that have at least one accurate retrieval. A retrieval is considered accurate if the 3D groundtruth from the retrieved pose satisfies the matching function (1) with $\kappa = 0.1$.

Baseline Approaches We compare Pr-VIPE with 2D-to-3D lifting models [32] and $L2$-VIPE. $L2$-VIPE outputs $L2$-normalized point embeddings, and is trained with the squared $L2$ distance kernel, similar to [50].

For fair comparison, we use the same backbone network architecture for all the models. Notably, this architecture [32] has been tuned for lifting tasks on H3.6M. Since the estimated 3D poses in camera coordinates are not view-invariant, we apply normalization and Procrustes alignment to align the estimated 3D poses between index and query for retrieval. In comparison, our embeddings do not require any alignment or other post-processing during retrieval.

For Pr-VIPE, we retrieve poses using nearest neighbors in the embedding space with respect to the sampled matching probability (7), which we refer to

Table 1: Comparison of cross-view pose retrieval results Hit@k (%) on H3.6M and 3DHP with chest-level cameras and all cameras. $*$ indicates that normalization and Procrustes alignment are performed on query-index pairs.

Dataset	H3.6M			3DHP (Chest)			3DHP (All)		
k	1	10	20	1	10	20	1	10	20
2D keypoints*	28.7	47.1	50.9	5.20	14.0	17.2	9.80	21.6	25.5
3D lifting*	69.0	89.7	92.7	24.9	54.4	62.4	24.6	53.2	61.3
$L2$-VIPE	73.5	94.2	96.6	23.8	56.7	66.5	18.7	46.3	55.7
$L2$-VIPE (w/ aug.)	70.4	91.8	94.5	24.9	55.4	63.6	23.7	53.0	61.4
Pr-VIPE	**76.2**	**95.6**	**97.7**	25.4	59.3	69.3	19.9	49.1	58.8
Pr-VIPE (w/ aug.)	73.7	93.9	96.3	**28.3**	**62.3**	**71.4**	**26.4**	**58.6**	**67.9**

as retrival confidence. We present the results on the VIPE models with and without camera augmentation. We applied similar camera augmentation to the lifting model, but did not see improvement in performance. We also show the results of pose retrieval using aligned 2D keypoints only. The poor performance of using input 2D keypoints for retrieval from different views confirms the fact that models must learn view invariance from inputs for this task.

We also compare with the image-based EpipolarPose model [26]. Please refer to the supplementary materials for the experiment details and results.

4.2.2 Quantitative Results From Table 1, we see that Pr-VIPE (with augmentation) outperforms all the baselines for H3.6M and 3DHP. The H3.6M results shown are on the hold-out set, and 3DHP is unseen during training, with more diverse poses and views. When we use all the cameras from 3DHP, we evaluate the generalization ability of models to new poses and new views. When we evaluate using only the 5 chest-level cameras from 3DHP, where the views are more similar to the training set in H3.6M, we mainly evaluate for generalization to new poses. When we evaluate using only the 5 chest-level cameras from 3DHP, the views are more similar to H3.6M, and generalization to new poses becomes more important. Our model is robust to the choice of β and the number of samples K (analysis in supplementary materials).

Table 1 shows that Pr-VIPE without camera augmentation is able to perform better than the baselines for H3.6M and 3DHP (chest-level cameras). This shows that Pr-VIPE is able to generalize as well as other baseline methods to new poses. However, for 3DHP (all cameras), the performance for Pr-VIPE without augmentation is worse compared with chest-level cameras. This observation indicates that when trained on chest-level cameras only, Pr-VIPE does not generalize as well to new views. The same results can be observed for $L2$-VIPE between chest-level and all cameras. In contrast, the 3D lifting models are able to generalize better to new views with the help of additional Procrustes alignment, which requires expensive SVD computation for every index-query pair.

We further apply camera augmentation to training the Pr-VIPE and the $L2$-VIPE model. Note that this step does not require camera parameters or additional groundtruth. The results in Table 1 on Pr-VIPE show that the aug-

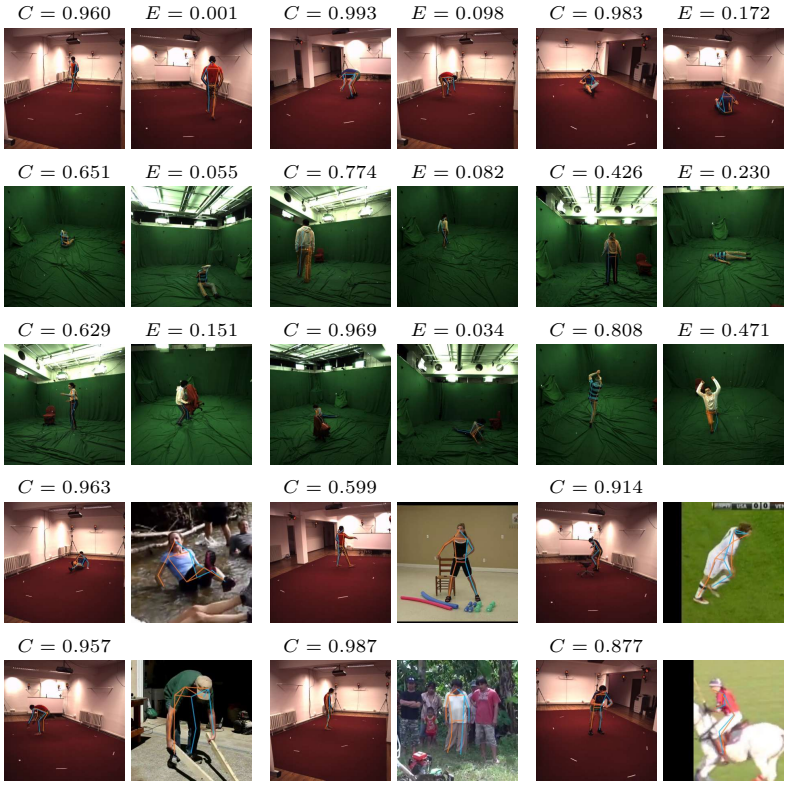


Fig. 3: Visualization of pose retrieval results. The first row is from H3.6M; the second and the third row are from 3DHP; the last two rows are using queries from H3.6M to retrieve from 2DHP. On each row, we show the query pose on the left for each image pair and the top-1 retrieval using the Pr-VIPE model (w/ aug.) on the right. We display retrieval confidences ("C") and top-1 NP-MPJPEs ("E", if 3D pose groundtruth is available).

mentation improves performance for 3DHP (all cameras) by 6% to 9%. This step also increases chest-level camera accuracy slightly. For $L2$-VIPE, we can observe a similar increase on all views. Camera augmentation reduces accuracy on H3.6M for both models. This is likely because augmentation reduces overfitting to the training camera views. By performing camera augmentation, Pr-VIPE is able to generalize better to new poses and new views.

4.2.3 Qualitative Results Fig. 3 shows qualitative retrieval results using Pr-VIPE. As shown in the first row, the retrieval confidence of the model is generally high for H3.6M. This indicates that the retrieved poses are close to their queries in the embedding space. Errors in 2D keypoint detection can lead to retrieval errors as shown by the rightmost pair. In the second and third rows, the retrieval confidence is lower for 3DHP. This is likely because there are new

poses and views unseen during training, which has the nearest neighbor slightly further away in the embedding space. We see that the model can generalize to new views as the images are taken at different camera elevations from H3.6M. Interestingly, the rightmost pair on row 2 shows that the model can retrieve poses with large differences in roll angle, which is not present in the training set. The rightmost pair on row 3 shows an example of a large NP-MPJPE error due to mis-detection of the left leg in the index pose.

We show qualitative results using queries from the H3.6M hold-out set to retrieve from 2DHP in the last two rows of Fig. 3. The results on these in-the-wild images indicate that as long as the 2D keypoint detector works reliably, our model is able to retrieve poses across views and subjects. More qualitative results are provided in the supplementary materials.

4.3 Downstream Tasks

We show that our pose embedding can be directly applied to pose-based downstream tasks using simple algorithms. We compare the performance of Pr-VIPE (**only trained on H3.6M, with no additional training**) on the Penn Action dataset against other approaches specifically trained for each task on the target dataset. In all the following experiments in this section, we compute our Pr-VIPE embeddings on single video frames and use the negative logarithm of the matching probability (7) as the distance between two frames. Then we apply temporal averaging within an atrous kernel of size 7 and rate 3 around the two center frames and use this averaged distance as the frame matching distance. Given the matching distance, we use standard dynamic time warping (DTW) algorithm to align two action sequences by minimizing the sum of frame matching distances. We further use the averaged frame matching distance from the alignment as the distance between two video sequences.

4.3.1 Action Recognition We evaluate our embeddings for action recognition using nearest neighbor search with the sequence distance described above. Provided person bounding boxes in each frame, we estimate 2D pose keypoints using [41]. On Penn Action, we use the standard train/test split [36]. Using all the testing videos as queries, we conduct two experiments: (1) we use all training videos as index to evaluate overall performance and compare with state-of-the-art methods, and (2) we use training videos only under one view as index and evaluate the effectiveness of our embeddings in terms of view-invariance. For this second experiment, actions with zero or only one sample under the index view are ignored, and accuracy is averaged over different views.

From Table 2 we can see that without any training on the target domain or using image context information, our embeddings can achieve highly competitive results on pose-based action classification, outperforming the existing best baseline that only uses pose input and even some other methods that rely on image context or optical flow. As shown in the last row in Table 2, our embeddings can be used to classify actions from different views using index samples from only one single view with relatively high accuracy, which further demonstrates the advantages of our view-invariant embeddings.

Table 2: Comparison of action recognition results on Penn Action.

Methods	Input RGB	Input Flow	Input Pose	Accuracy (%)
Nie *et al.* [36]	✓		✓	85.5
Iqbal *et al.* [18]			✓	79.0
Cao *et al.* [5]		✓	✓	95.3
	✓	✓		98.1
Du *et al.* [10]	✓	✓	✓	97.4
Liu *et al.* [30]	✓		✓	91.4
Luvizon *et al.* [31]	✓		✓	98.7
Ours			✓	97.5
Ours (1-view index)			✓	92.1

Table 3: Comparison of video alignment results on Penn Action.

Methods	Kendall's Tau
SaL [34]	0.6336
TCN [51]	0.7353
TCC [11]	0.7328
TCC + SaL [11]	0.7286
TCC + TCN [11]	0.7672
Ours	0.7476
Ours (same-view only)	0.7521
Ours (different-view only)	0.7607

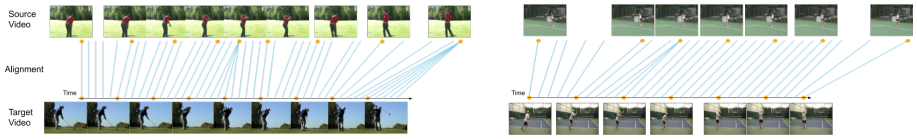


Fig. 4: Video alignment results using Pr-VIPE. The orange dots correspond to the visualized frames, and the blue line segments illustrate the frame alignment.

4.3.2 Video Alignment Our embeddings can be used to align human action videos from different views using DTW algorithm as described earlier in Section 4.3. We measure the alignment quality of our embeddings quantitatively using Kendall's Tau [11], which reflects how well an embedding model can be applied to align unseen sequences if we use nearest neighbor in the embedding space to match frames for video pairs. A value of 1 corresponds to perfect alignment. We also test the view-invariant properties of our embeddings by evaluating Kendall's Tau on aligning videos pairs from the same view, and aligning pairs with different views.

In Table 3, we compare our results with other video embedding baselines that are trained for the alignment task on Penn Action, from which we observe that Pr-VIPE performs better than all the method that use a single type of loss. While Pr-VIPE is slightly worse than the combined TCC+TCN loss, our embeddings are able to achieve this without being explicitly trained for this task or taking advantage of image context. In the last two rows of Table 3, we show the results from evaluating video pairs only from the same or different views. We can see that our embedding achieves consistently high performance regardless of whether the aligned video pair is from the same or different views, which demonstrate its view-invariant property. In Fig. 4, we show action video synchronization results from different views using Pr-VIPE. We provide more synchronized videos for all actions in the supplementary materials.

4.4 Ablation Study

Point vs. Probabilistic Embeddings We compare VIPE point embedding formulation with Pr-VIPE. When trained on detected keypoints, the Hit@1 for

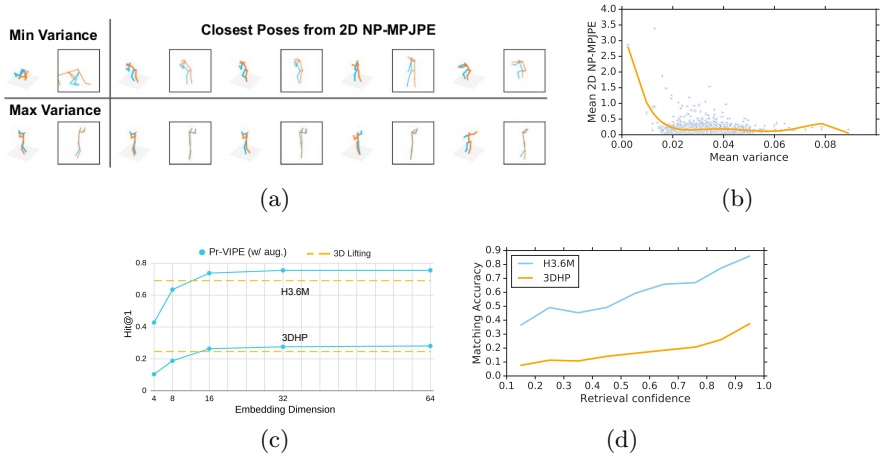


Fig. 5: Ablation study: (a) Top retrievals by 2D NP-MPJPE from the H3.6M hold-out subset for queries with largest and smallest variance. 2D poses are shown in the boxes. (b) Relationship between embedding variance and 2D NP-MPJPE to top-10 nearest 2D pose neighbors from the H3.6M hold-out subset. The orange curve represents the best fitting 5th degree polynomial. (c) Comparison of Hit@1 with different embedding dimensions. The 3D lifting baseline predictss 39 dimensions. (d) Relationship between retrieval confidence and matching accuracy.

VIPE and Pr-VIPE are 75.4% and 76.2% on H3.6M, and 19.7% and 20.0% on 3DHP, respectively. When we add camera augmentation, the Hit@1 for VIPE and Pr-VIPE are 73.8% and 73.7% on H3.6M, and 26.1% and 26.5% on 3DHP, respectively. Despite the similar retrieval accuracies, Pr-VIPE is generally more accurate and, more importantly, has additional desirable properties in that the variance can model 2D input ambiguity as to be discussed next.

A 2D pose is ambiguous if there are similar 2D poses that can be projected from very different poses in 3D. To measure this, we compute the average 2D NP-MPJPE between a 2D pose and its top-10 nearest neighbors in terms of 2D NP-MPJPE. To ensure the 3D poses are different, we sample 1200 poses from H3.6M hold-out set with a minimum gap of 0.1 3D NP-MPJPE. If a 2D pose has small 2D NP-MPJPE to its neighbors, it means there are many similar 2D poses corresponding to different 3D poses and so the 2D pose is ambiguous.

Fig. 5a shows that the 2D pose with the largest variance is ambiguous as it has similar 2D poses in H3.6M with different 3D poses. In contrast, we see that the closest 2D poses corresponding to the smallest variance pose on the first row of Fig. 5a are clearly different. Fig. 5b further shows that as the average variance increases, the 2D NP-MPJPE between similar poses generally decreases, which means that 2D poses with larger variances are more ambiguous.

Embedding Dimensions Fig. 5c demonstrates the effect of embedding dimensions on H3.6M and 3DHP. The lifting model lifts 13 2D keypoints to 3D, and therefore has a constant output dimension of 39. We see that Pr-VIPE (with augmentation) is able to achieve a higher accuracy than lifting at 16 dimensions.

Additionally, we can increase the number of embedding dimensions to 32, which increases accuracy of Pr-VIPE from 73.7% to 75.5%.

Retrieval Confidence In order to validate the retrieval confidence values, we randomly sample 100 queries along with their top-5 retrievals (using Pr-VIPE retrieval confidence) from each query-index camera pair. This procedure forms 6000 query-retrieval sample pairs for H3.6M (4 views, 12 camera pairs) and 55000 for 3DHP (11 views, 110 camera pairs), which we bin by their retrieval confidences. Fig. 5d shows the matching accuracy for each confidence bin. We can see that the accuracy positively correlates with the confidence values, which suggest our retrieval confidence is a valid indicator to model performance.

What if 2D keypoint detectors were perfect? We repeat our pose retrieval experiments using groundtruth 2D keypoints to simulate a perfect 2D keypoint detector on H3.6M and 3DHP. All experiments use the 4 views from H3.6M for training following the standard protocol. For the baseline lifting model in camera frame, we achieve 89.9% Hit@1 on H3.6M, 48.2% on 3DHP (all), and 48.8% on 3DHP (chest). For Pr-VIPE, we achieve 97.5% Hit@1 on H3.6M, 44.3% on 3DHP (all), and 66.4% on 3DHP (chest). These results follow the same trend as using detected keypoints inputs in Table 1. Comparing the results with using detected keypoints, the large improvement in performance using groundtruth keypoints suggests that a considerable fraction of error in our model is due to imperfect 2D keypoint detections. Please refer to the supplementary materials for more ablation studies and embedding space visualization.

5 Conclusion

We introduce Pr-VIPE, an approach to learning probabilistic view-invariant embeddings from 2D pose keypoints. By working with 2D keypoints, we can use camera augmentation to improve model generalization to unseen views. We also demonstrate that our probabilistic embedding learns to capture input ambiguity. Pr-VIPE has a simple architecture and can be potentially applied to object and hand poses. For cross-view pose retrieval, 3D pose estimation models require expensive rigid alignment between query-index pair, while our embeddings can be applied to compare similarities in simple Euclidean space. In addition, we demonstrated the effectiveness of our embeddings on downstream tasks for action recognition and video alignment. Our embedding focuses on a single person, and for future work, we will investigate extending it to multiple people and robust models that can handle missing keypoints from input.

Acknowledgment

We thank Yuxiao Wang, Debidatta Dwibedi, and Liangzhe Yuan from Google Research, Long Zhao from Rutgers University, and Xiao Zhang from University of Chicago for helpful discussions. We appreciate the support of Pietro Perona, Yisong Yue, and the Computational Vision Lab at Caltech for making this collaboration possible. The author Jennifer J. Sun is supported by NSERC (funding number PGSD3-532647-2019) and Caltech.

References

1. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: CVPR (2015)
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
3. Bojchevski, A., Günnemann, S.: Deep Gaussian embedding of graphs: Unsupervised inductive learning via ranking (2017)
4. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: NeurIPS (1994)
5. Cao, C., Zhang, Y., Zhang, C., Lu, H.: Body joint guided 3-D deep convolutional descriptors for action recognition. IEEE Transactions on Cybernetics **48**(3), 1095–1108 (2017)
6. Chen, C.H., Ramanan, D.: 3D human pose estimation = 2D pose estimation + matching. In: CVPR (2017)
7. Chen, C.H., Tyagi, A., Agrawal, A., Drover, D., Stojanov, S., Rehg, J.M.: Unsupervised 3D pose estimation with geometric self-supervision. In: CVPR (2019)
8. Chu, R., Sun, Y., Li, Y., Liu, Z., Zhang, C., Wei, Y.: Vehicle re-identification with viewpoint-aware metric learning. In: ICCV (2019)
9. Drover, D., Chen, C.H., Agrawal, A., Tyagi, A., Phuoc Huynh, C.: Can 3D pose be learned from 2D projections alone? In: ECCV (2018)
10. Du, W., Wang, Y., Qiao, Y.: RPAN: An end-to-end recurrent pose-attention network for action recognition in videos. In: ICCV (2017)
11. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: CVPR (2019)
12. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR (2006)
13. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv:1703.07737 (2017)
14. Ho, C.H., Morgado, P., Persekian, A., Vasconcelos, N.: PIEs: Pose invariant embeddings. In: CVPR. pp. 12377–12386 (2019)
15. Hu, W., Zhu, S.C.: Learning a probabilistic model mixing 3D and 2D primitives for view invariant object recognition. In: CVPR (2010)
16. Huang, C., Loy, C.C., Tang, X.: Local similarity-aware deep feature embedding. In: NeurIPS (2016)
17. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE TPAMI (2013)
18. Iqbal, U., Garbade, M., Gall, J.: Pose for action-action for pose. In: FG (2017)
19. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Mining on manifolds: Metric learning without labels. In: CVPR (2018)
20. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. arXiv:1905.05754 (2019)
21. Jammalamadaka, N., Zisserman, A., Eichner, M., Ferrari, V., Jawahar, C.: Video retrieval by mimicking poses. In: ACM ICMR (2012)
22. Ji, X., Liu, H.: Advances in view-invariant human motion analysis: A review. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **40**(1), 13–24 (2009)
23. Ji, X., Liu, H., Li, Y., Brown, D.: Visual-based view-invariant human motion analysis: A review. In: International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (2008)

24. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NeurIPS (2017)
25. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2014)
26. Kocabas, M., Karagoz, S., Akbas, E.: Self-supervised learning of 3D human pose using multi-view geometry (2019)
27. LeCun, Y., Huang, F.J., Bottou, L., et al.: Learning methods for generic object recognition with invariance to pose and lighting. In: CVPR (2004)
28. Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.: Unsupervised learning of view-invariant action representations. In: NeurIPS (2018)
29. Liu, J., Akhtar, N., Ajmal, M.: Viewpoint invariant action recognition using RGB-D videos. IEEE Access **6**, 70061–70071 (2018)
30. Liu, M., Yuan, J.: Recognizing human actions as the evolution of pose estimation maps. In: CVPR (2018)
31. Luvizon, D.C., Tabia, H., Picard, D.: Multi-task deep learning for real-time 3D human pose estimation and action recognition. arXiv:1912.08077 (2019)
32. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: ICCV (2017)
33. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3D human pose estimation in the wild using improved CNN supervision. In: 3DV (2017)
34. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: Unsupervised learning using temporal order verification. In: ECCV (2016)
35. Mori, G., Pantofaru, C., Kothari, N., Leung, T., Toderici, G., Toshev, A., Yang, W.: Pose embeddings: A deep architecture for learning to match human poses. arXiv:1507.00302 (2015)
36. Nie, B.X., Xiong, C., Zhu, S.C.: Joint action recognition and pose estimation from video. In: CVPR (2015)
37. Oh, S.J., Murphy, K., Pan, J., Roth, J., Schroff, F., Gallagher, A.: Modeling uncertainty with hedged instance embedding. ICLR (2019)
38. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: CVPR (2016)
39. Ong, E.J., Micilotta, A.S., Bowden, R., Hilton, A.: Viewpoint invariant exemplar-based 3D human tracking. CVIU (2006)
40. Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: ECCV (2018)
41. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: CVPR (2017)
42. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC (2015)
43. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: CVPR (2019)
44. Qiu, H., Wang, C., Wang, J., Wang, N., Zeng, W.: Cross View Fusion for 3D Human Pose Estimation. arXiv:1909.01203 (2019)
45. Rao, C., Shah, M.: View-invariance in action recognition. In: CVPR (2001)
46. Rayat Imtiaz Hossain, M., Little, J.J.: Exploiting temporal information for 3D human pose estimation. In: ECCV (2018)
47. Rhodin, H., Constantin, V., Katircioglu, I., Salzmann, M., Fua, P.: Neural scene decomposition for multi-person motion capture. In: CVPR (2019)

48. Rhodin, H., Salzmann, M., Fua, P.: Unsupervised geometry-aware representation for 3D human pose estimation. In: ECCV (2018)
49. Rhodin, H., Spörri, J., Katircioglu, I., Constantin, V., Meyer, F., Müller, E., Salzmann, M., Fua, P.: Learning monocular 3D human pose estimation from multi-view images. In: CVPR (2018)
50. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: CVPR (2015)
51. Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., Brain, G.: Time-contrastive networks: Self-supervised learning from video. In: ICRA (2018)
52. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: ECCV (2018)
53. Tekin, B., Márquez-Neila, P., Salzmann, M., Fua, P.: Learning to fuse 2D and 3D image cues for monocular body pose estimation. In: ICCV (2017)
54. Tome, D., Toso, M., Agapito, L., Russell, C.: Rethinking pose in 3D: Multi-stage refinement and recovery for markerless motion capture. In: 3DV (2018)
55. Vilnis, L., McCallum, A.: Word representations via gaussian embedding. arXiv:1412.6623 (2014)
56. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: CVPR (2014)
57. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3D pose estimation. In: CVPR (2015)
58. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: ICCV (2017)
59. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: CVPRW (2012)
60. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: ICCV (2013)
61. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose invariant embedding for deep person re-identification. IEEE TIP (2019)
62. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3D human pose estimation in the wild: A weakly-supervised approach. In: ICCV (2017)