Learning View-Disentangled Human Pose Representation by Contrastive Cross-View Mutual Information Maximization

Long Zhao[1,*]   Yuxiao Wang[2]   Jiaping Zhao[2]   Liangzhe Yuan[2]   Jennifer J. Sun[3]
Florian Schroff[2]   Hartwig Adam[2]   Xi Peng[4]   Dimitris Metaxas[1]   Ting Liu[2]

[1]Rutgers University   [2]Google Research   [3]Caltech   [4]University of Delaware

Abstract

We introduce a novel representation learning method to disentangle pose-dependent as well as view-dependent factors from 2D human poses. The method trains a network using cross-view mutual information maximization (CV-MIM) which maximizes mutual information of the same pose performed from different viewpoints in a contrastive learning manner. We further propose two regularization terms to ensure disentanglement and smoothness of the learned representations. The resulting pose representations can be used for cross-view action recognition.

To evaluate the power of the learned representations, in addition to the conventional fully-supervised action recognition settings, we introduce a novel task called single-shot cross-view action recognition. This task trains models with actions from only one single viewpoint while models are evaluated on poses captured from all possible viewpoints. We evaluate the learned representations on standard benchmarks for action recognition, and show that (i) CV-MIM performs competitively compared with the state-of-the-art models in the fully-supervised scenarios; (ii) CV-MIM outperforms other competing methods by a large margin in the single-shot cross-view setting; (iii) and the learned representations can significantly boost the performance when reducing the amount of supervised training data. Our code is made publicly available at https://github.com/google-research/google-research/tree/master/poem.

1. Introduction

Understanding human poses and actions is a fundamental problem in computer vision due to its broad applications in the real world, such as video content analysis, intelligent photography, AR/VR techniques, and human-computer interface. Recently, remarkable improvements have been

---

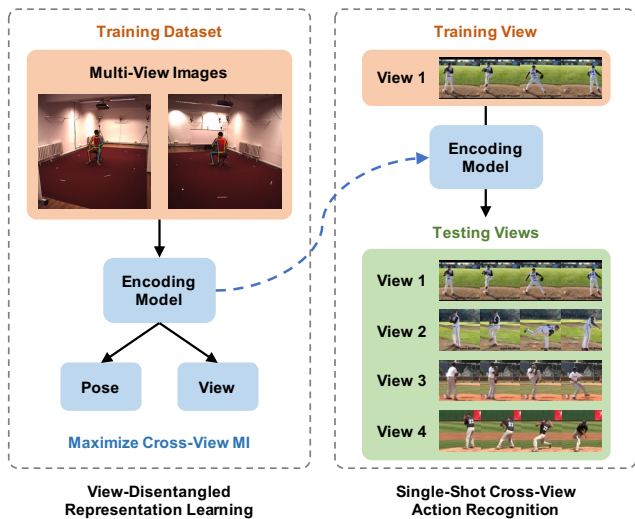*This work was done while the author was a research intern at Google.

Figure 1. **Left:** We propose to learn view-disentangled representation for human poses by maximizing cross-view mutual information. **Right:** The learned representation can be applied to downstream tasks such as single-shot cross-view action recognition.

achieved with deep learning approaches [9, 20, 28, 54, 55]. However, these data-driven approaches are usually vulnerable to changes of viewpoints. In particular, testing-time unseen viewpoints often lead to significant degradation in recognition performance [47].

To mitigate this issue, methods for cross-view action recognition [47] have been proposed, where models are trained with a set of actions captured from different viewpoints simultaneously so that they can be applied to novel views unseen from training at testing time. Previous studies usually require extensive supervision from multiple views to learn view-invariant features [46, 47] or transferable representations [21, 23] for action recognition. Collecting labeled action data at scale from multiple views can be expensive and challenging in the wild due to potential limitation of camera placement, scene and actor setup, *etc.*

We address this challenge by proposing a novel view-

disentangled representation learning approach. To train the representation model, we only require pairs of 2D poses captured from different viewpoints without additional task-relevant supervision, which are widely available in standard multi-view human action datasets [17, 38]. Our target is to disentangle pose-dependent (view-invariant) and view-dependent representations from 2D poses, which has not been well-explored in existing works.

To achieve this, we train a representation-learning function, *i.e.*, an encoder, following the Mutual Information (MI) maximization principle [4]. Specifically, in order to fulfill the view-disentanglement constraint, we propose to maximize the cross-view MI, *i.e.*, the dependency between learned representations of the same pose from different views. In addition, we theoretically motivate two regularization terms that encourage disentanglement and smoothness of the learned representation to further improve its power. Our objective is optimized in a contrastive manner based on recent advances made in MI estimation [3, 7, 15, 30, 31]. Compared to approaches based on cross reconstruction [29], the proposed approach yields stronger representative powers by using negative training pairs which provide an additional source of supervision [6, 15].

We show that the resulting pose representation can be used for action recognition in a fully-supervised setting. To further demonstrate its view-disentangled property, we introduce a novel and more challenging task, namely, single-shot cross-view action recognition. In this setting, recognition models are trained with 2D poses from one single view but expected to generalize to unseen views at testing time. This setting is highly practical for real-world applications: it only requires collecting training data from a fixed camera view, and the resulting recognition model can be applied to various difference views. Note that success in this task requires not only discriminative but also view-invariant representations for 2D poses. Fig. 1 summarizes our representation learning framework and its application to single-shot cross-view action recognition downstream task.

To sum up, our main contributions of this work include: (i) a novel objective to learn view-disentangled representation for 2D human poses by maximizing cross-view MI; (ii) two regularization techniques to guarantee disentanglement and smoothness of the learned representation; (iii) a newly proposed task called single-shot cross-view action recognition that can be used for evaluating view-invariant representation for human poses. We evaluate the proposed method on standard benchmarks for action recognition under scenarios of full-supervision, single-shot cross-view setting, and supervision with limited data. Experimental results show that our approach is comparable to the state of the art in the fully-supervised setting, while it can consistently and statistically significantly outperform competing methods under the other two scenarios.

2. Related Work

Representation Learning. There has been a lot of recent progress on learning representations for visual and temporal data [2, 6, 13, 12, 15, 30]. These representations are often trained using unsupervised or self-supervised approaches. In particular, approaches based on contrastive learning (contrasting positive pairs with negative pairs) has been shown to be effective at learning visual representations [6, 12, 30, 44]. Recent works have further investigated contrastive training objectives based on MI maximization, such as maximizing MI between different augmented "views" of the same image [2] and between local and global features [15, 30].

Our work aims to learn a representation on 2D poses instead of images and we apply contrastive learning to maximize MI of pose representations across camera views. Previous works on representation learning for 2D poses [43] have focused on studying the view-invariance property with triplet loss [37]. In contrast, we differ in our goal, *i.e.*, disentangling pose-dependent and view-dependent factors, and our approach, *i.e.*, contrastive loss with MI maximization across camera views.

View Disentanglement. When depicted in 2D space, human poses can differ in appearance due to changes in pose and changes in viewpoints. The ability to disentangle pose-dependent and view-dependent factors from human poses as well as objects are useful for a variety of downstream tasks, including video alignment [11], human re-identification [56], action recognition [29], and object classification [16]. Here, we explore view-disentanglement for pose-based action recognition.

Some studies focus on learning view-invariant representations for human poses [43] and objects [16]. The learned representation space in these works is pose-dependent, and not view-dependent. Other works [25, 29, 33, 35, 36, 41] also learns disentangled representations for human poses and images. In particular, [35, 36] learns a 3D geometry-aware representation space for pose images, where rotation matrices can be applied to the representation to generate images from new views; [25, 41] disentangle shape and appearance using generative modeling on images, and does not explicitly disentangle viewpoints. Closest to our work, [29] also learns to disentangle poses and views on human poses. Our work differs in that we embed 2D poses, which can be extracted from images without using camera parameters, while their method relies on ground truth 3D poses as input. Additionally, instead of cross reconstruction, our approach is based on contrastive loss and MI maximization.

Cross-View Learning for Poses. There have been great progress in cross-view action recognition with RGB videos. In the multi-view environment, many research works aim to address the issue of view invariance motivated by the availability of different modalities such as pose and depth. In
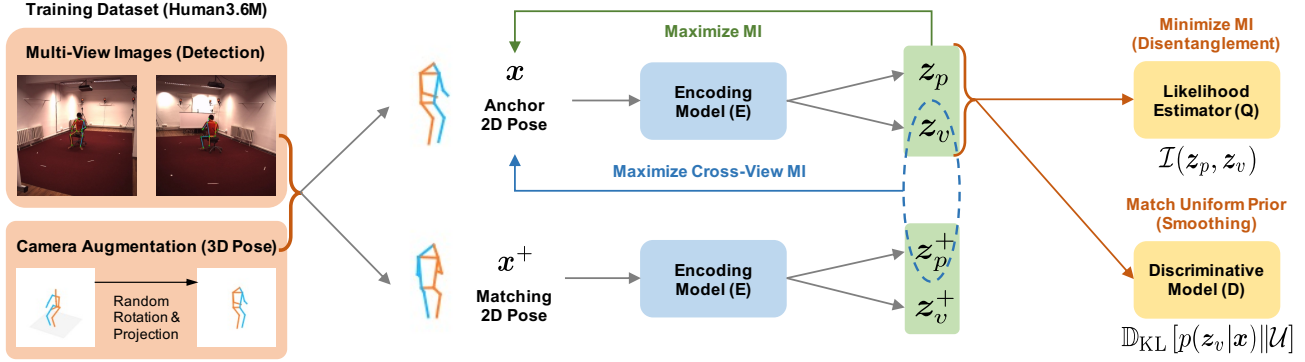
Figure 2. Overview of CV-MIM model training pipeline. Our model takes pairs of multi-view 2D poses detected from images or camera augmentation (optional) and produces pose representations $\boldsymbol{z}_p$ and view representations $\boldsymbol{z}_v$. E, Q and D are optimized alternatively.

this stream of research, most works use multiple modalities, *e.g.*, RGB [22, 45, 52], depth [22, 34, 45], RGB+D [39], or skeleton data [46, 50, 51]. Another stream of research works [47, 21] are interested in learning models by using poses from different camera views for cross-view action recognition. Our work belongs to this stream.

Multi-view pose information has been used for learning image representations for 3D pose estimation [36, 35] as well. In these works, the representation is trained to reconstruct images of poses from different camera views. While we also leverage multi-view pose data, our method is based on cross-view MI maximization. It can utilize negative samples during optimization for representation learning.

3. Approach

We begin by summarizing the concept of mutual information and, along the way, introduce the notations. Mutual Information (MI) is a fundamental measurement to quantify the relationship between random variables. Formally, it measures the dependence of two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\mathcal{I}(\boldsymbol{x}; \boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{x}, \boldsymbol{y})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x}) p(\boldsymbol{y})} \right], \qquad (1)$$

where $p(\boldsymbol{x}, \boldsymbol{y})$ is the joint probability distribution, while $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ are their marginals. In the context of self-supervised representation learning, mutual information can act as a measure of true dependence between observed data samples and learned representations. The objective is to maximize Eq. (1) so that the learned representations retain the most information about the underlying data [4, 15, 31].

This work extends MI maximization principle to view-disentangled representation learning for human poses where view-dependent and pose-dependent representations are learned concurrently for an input 2D pose. As we will show, this objective can be obtained by maximizing cross-view MI. An overview of our approach is presented in Fig. 2.

3.1. Cross-View Mutual Information Maximization

To set the stage, we let $\boldsymbol{x}^i \in \mathbb{R}^{2 \times N}$ denote the given 2D pose from the i-th view, where N is the number of keypoints for representing the pose. We are interested in learning an encoding network E producing a view representation $\boldsymbol{z}_v^i \in \mathbb{R}^d$ and a pose representation $\boldsymbol{z}_p^i \in \mathbb{R}^d$ from the input $\boldsymbol{x}^i$, while $\boldsymbol{z}_p^i$ and $\boldsymbol{z}_v^i$ are expected to be disentangled (mutually excluded). We define $(\boldsymbol{z}_p^j, \boldsymbol{z}_v^i)$ as the cross-view representation of a given 2D pose $\boldsymbol{x}^i$ from the j-th view, and it captures the amount of pose information that can be maintained from a different viewpoint. Then we can obtain our objective for view-disentangled representation learning:

$$\max \left[\sum_i \underbrace{\mathcal{I}(\boldsymbol{x}^i; \boldsymbol{z}_p^i, \boldsymbol{z}_v^i)}_{\text{MI}} + \sum_{i \neq j} \underbrace{\mathcal{I}(\boldsymbol{x}^i; \boldsymbol{z}_p^j, \boldsymbol{z}_v^i)}_{\text{Cross-View MI}} \right], \quad (2)$$

where the first term is the conventional MI-based representation objective, and the second term which is proposed in this work maximizes the MI between the input 2D pose and its cross-view representations. Next, we show how this objective can be further simplified.

As the optimal view and pose representations are disentangled, we assume that $\boldsymbol{z}_p$ and $\boldsymbol{z}_v$ are simultaneously independent and conditionally independent, *i.e.*, $p(\boldsymbol{z}_p, \boldsymbol{z}_v) = p(\boldsymbol{z}_p) p(\boldsymbol{z}_v)$ and $p(\boldsymbol{z}_p, \boldsymbol{z}_v | \boldsymbol{x}) = p(\boldsymbol{z}_p | \boldsymbol{x}) p(\boldsymbol{z}_v | \boldsymbol{x})$. Based on this assumption, it can be easily shown that,

$$\mathcal{I}(\boldsymbol{x}^i; \boldsymbol{z}_p^j, \boldsymbol{z}_v^i) = \mathcal{I}(\boldsymbol{x}^i; \boldsymbol{z}_p^j) + \mathcal{I}(\boldsymbol{x}^i; \boldsymbol{z}_v^i). \qquad (3)$$

Then, by the Data Processing Inequality [8], we have that,

$$\begin{aligned}
\mathcal{I}(\boldsymbol{x}^i; \boldsymbol{z}_p^j, \boldsymbol{z}_v^i) &\geqslant \mathcal{I}(\boldsymbol{z}_p^i, \boldsymbol{z}_v^i; \boldsymbol{z}_p^j) + \mathcal{I}(\boldsymbol{z}_p^i, \boldsymbol{z}_v^i; \boldsymbol{z}_v^i) \\
&= \mathcal{I}(\boldsymbol{z}_p^i; \boldsymbol{z}_p^j) + \mathcal{I}(\boldsymbol{z}_v^i; \boldsymbol{z}_v^i) \\
&= \mathcal{I}(\boldsymbol{z}_p^i; \boldsymbol{z}_p^j) + \mathcal{H}(\boldsymbol{z}_v^i) \\
&\geqslant \mathcal{I}(\boldsymbol{z}_p^i; \boldsymbol{z}_p^j),
\end{aligned} \qquad (4)$$

where $\mathcal{H}$ is the Shannon entropy. The equality in the second line is achieved since $\boldsymbol{z}_p$ and $\boldsymbol{z}_v$ are independent, and the last inequality holds due to the non-negativity of entropy.

After combing Eq. (4) with Eq. (2), we achieve a relaxed formulation of cross-view MI maximization:

$$\max \left[\sum_i \mathcal{I}(\boldsymbol{x}^i; \boldsymbol{z}_p^i, \boldsymbol{z}_v^i) + \sum_{i \neq j} \mathcal{I}(\boldsymbol{z}_p^i; \boldsymbol{z}_p^j) \right]. \quad (5)$$

The above objective is a lower bound of Eq. (2). Intuitively, the second term aims to maximize MI of the same pose but performed from different viewpoints.

Both terms in Eq. (5) can be optimized by MI estimators [3, 15, 30, 31] which estimate a lower-bound of MI by training a classifier in a contrastive learning objective, *i.e.*, it distinguishes between samples coming from the joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$ and the product of marginals $p(\boldsymbol{x})p(\boldsymbol{z})$ of the input pose $\boldsymbol{x}$ and target representation $\boldsymbol{z}$ encoded by E. We use the Jensen-Shannon MI estimator [15] to maximize Eq. (5) since it achieves a good balance between computational efficiency and performance. Given the representation $\boldsymbol{z}'$ which is a negative match of the input $\boldsymbol{x}$, maximizing $\mathcal{I}(\boldsymbol{x}; \boldsymbol{z})$ is equivalent to minimizing the following loss:

$$\min_E \mathcal{L}_{\text{MI}}(\boldsymbol{x}; \boldsymbol{z}) = \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{z}) \sim p(\boldsymbol{x}, \boldsymbol{z})}[\rho(-f(\boldsymbol{x}, \boldsymbol{z}))]$$
$$- \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{z}') \sim p(\boldsymbol{x})p(\boldsymbol{z})}[-\rho(f(\boldsymbol{x}, \boldsymbol{z}'))], \quad (6)$$

where $\rho(x) = \log(1 + e^x)$ denotes the softplus activation, and f is a discriminator function modeled by a network.

3.2. Representation Disentanglement

The major assumption we made when deriving Eq. (3) is that $\boldsymbol{z}_v$ (view representation) and $\boldsymbol{z}_p$ (pose representation) are disentangled during optimization, *i.e.*, they are simultaneously and conditionally independent. Therefore, we introduce a regularization term $\mathcal{L}_{\text{inter}}$ to guarantee disentanglement between $\boldsymbol{z}_v$ and $\boldsymbol{z}_p$ based on their MI. By minimizing $\mathcal{L}_{\text{inter}}$, we encourage the information in these two random variables are mutually exclusive.

However, lower-bound MI estimators are inapplicable to disentanglement because they are inconsistent to MI minimization tasks. Hence, we leverage the contrastive log-ratio upper-bound MI estimator [7] which estimates the probability log-ratio between the conditional log-likelihood of positive sample pair $\log p(\boldsymbol{z}_v|\boldsymbol{z}_p)$ and negative sample pair $\log p(\boldsymbol{z}_v'|\boldsymbol{z}_p)$. Unfortunately, the conditional relation between $\boldsymbol{z}_v$ and $\boldsymbol{z}_p$ is unavailable in our case. To address this issue, we use a variational distribution $q(\boldsymbol{z}_v|\boldsymbol{z}_p)$ which is predicted by a neural network Q to approximate $p(\boldsymbol{z}_v|\boldsymbol{z}_p)$. After combining all these together, we reach the following objective function for the encoder E:

$$\min_E \mathcal{L}_{\text{inter}}(\boldsymbol{z}_p; \boldsymbol{z}_v) = \mathbb{E}_{(\boldsymbol{z}_p, \boldsymbol{z}_v) \sim p(\boldsymbol{z}_p, \boldsymbol{z}_v)}[\log q(\boldsymbol{z}_v|\boldsymbol{z}_p)]$$
$$- \mathbb{E}_{(\boldsymbol{z}_p, \boldsymbol{z}_v') \sim p(\boldsymbol{z}_p)p(\boldsymbol{z}_v)}[\log q(\boldsymbol{z}_v'|\boldsymbol{z}_p)]. \quad (7)$$

While at the same time, Q is trained to minimize the KL-divergence between the true conditional probability distribution $p(\boldsymbol{z}_v|\boldsymbol{z}_p)$ and variational one $q(\boldsymbol{z}_v|\boldsymbol{z}_p)$:

$$\min_Q \mathcal{L}_{\text{KL}}(\boldsymbol{z}_p, \boldsymbol{z}_v) = \mathbb{D}_{\text{KL}}[q(\boldsymbol{z}_v|\boldsymbol{z}_p) \| p(\boldsymbol{z}_v|\boldsymbol{z}_p)]. \quad (8)$$

For simplicity, we assume $q(\boldsymbol{z}_v|\boldsymbol{z}_p)$ follows a Gaussian distribution in this work, and then Eq. (8) can be efficiently solved by maximum likelihood estimation.

3.3. Representation Smoothing

From Eq. (4) we can see that maximizing the entropy of the view representation $\mathcal{H}(\boldsymbol{z}_v)$ is also desirable during optimization. However, this term is intractable due to the high dimensionality of the representation space. As we are not concerned with its precise value, we present an alternative maximization strategy from the aspect of prior matching.

Given a bounded interval $[a, b]$, entropy is maximised when the probability distribution is uniform. By this observation and $\mathcal{H}(\boldsymbol{z}_v) \geqslant \mathcal{H}(\boldsymbol{z}_v|\boldsymbol{x})$, we impose the maximum-entropy constraint onto learned representations by implicitly training the encoder E so that the push-forward distribution $p(\boldsymbol{z}_v|\boldsymbol{x})$ matches a uniform prior $\mathcal{U}(a, b)$:

$$\min \{ \mathbb{D}_{\text{KL}}[p(\boldsymbol{z}_v|\boldsymbol{x}) \| \mathcal{U}(a, b)] \}. \quad (9)$$

This is achieved by training a discriminator D to estimate the divergence in Eq. (9), and then training the encoder E to minimize this estimation. They play the minimax game:

$$\min_E \max_D \mathcal{L}_{\text{prior}}(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{U}(a, b)}[\log D(\boldsymbol{z})]$$
$$+ \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{x})}[\log(1 - D(\boldsymbol{z}))]. \quad (10)$$

To keep it simple, we optimize this loss term on $[0, 1]$ which is done by setting the prior to $\mathcal{U}(0, 1)$ and re-scaling representations via a sigmoid activation. In practice, we match both view and pose representations to this prior since it also benefits the regularization of pose dimensions.

Intuitively, the loss term $\mathcal{L}_{\text{prior}}$ ensures the learned representations to be smooth [48], as we do not assume any special prior on human poses and camera views. Compared with previous works [1, 14, 15, 19, 43] that also target representation regularization, our approach provides a more intuitive motivation in favor of uniform prior over other common priors, *e.g.*, a Gaussian distribution.

3.4. Full Objective

All three objectives, *i.e.*, MI maximization, representation disentanglement and smoothing (prior matching), can be used together, and doing so we arrive at our full objective for *Cross-View Mutual Information Maximization* (CV-MIM). We let $\boldsymbol{x}^+$ represent a positive match of the input pose $\boldsymbol{x}$ which shares the same action but performed from another view, and $\boldsymbol{z}_p^+$ be its pose representation, then the complete objective is defined as:

$$
\begin{aligned}
\min_E \big[\mathcal{L}_{\mathrm{MI}}(\boldsymbol{x}; \boldsymbol{z}_p \odot \boldsymbol{z}_v) + \lambda_1 \mathcal{L}_{\mathrm{MI}}(\boldsymbol{z}_p; \boldsymbol{z}_p^+) \big] & \\
+ \min_E \lambda_2 \mathcal{L}_{\mathrm{inter}}(\boldsymbol{z}_p; \boldsymbol{z}_v) + \min_Q \mathcal{L}_{\mathrm{KL}}(\boldsymbol{z}_p, \boldsymbol{z}_v) & \\
+ \min_E \max_D \lambda_3 \mathcal{L}_{\mathrm{prior}}(\boldsymbol{z}_p \oplus \boldsymbol{z}_v), \quad (11) &
\end{aligned}
$$

where λ_1, λ_2, and λ_3 are positive parameters that balance the magnitude of each term; $\odot$ is a pre-defined fusion operation which combines pose and view representations; and $\oplus$ denotes concatenation. Note that E, Q and D are optimized in an alternative way during network training.

Discussion. It is worth discussing two important properties of our formulation. First, our approach differs from cross-reconstruction based methods [29, 33, 35, 36] from two perspectives: (i) we do not explicitly perform reconstruction of the input in the objective, which is proven to be a lower bound of MI [15]; (ii) our objective trains the model in a contrastive manner, where negative sample pairs are involved to provide additional supervisions and thus manage to improve the power of representation learning.

Second, in addition to Eq. (11), an alternative way of cross-view MI maximization is to optimize Eq. (2) directly through lower-bound MI estimators. However, we find this alternate leads to significant performance degradation in the experiments. This is due to the fact that lower-bound MI estimators are not accurate estimations of MI which suffer from high variance [42]. Instead, our formulation is able to address this drawback by decomposing the single objective into multiple simpler criteria.

4. Experiments

4.1. Datasets

Human3.6M. Ionescu *et al*. [17] built the in-lab dataset with synchronized multi-view images and 3D poses. We follow the standard protocol in the literature and use all four camera views of subjects S1, S5, S6, S7, and S8 for model training. Note that this dataset is only used for learning pose representation in this work, *i.e.*, training the encoder, where we do not use any action labels.

We experiment on the following two datasets for action recognition in the fully-supervised scenario where training sets include all views, and the single-shot cross-view setting where actions from only one view are used for training.

Penn Action. The Penn Action dataset [53] consists of 2,326 video sequences of 15 action categories captured from four different views. We follow the official training/testing split [53] and [28] to remove the action of playing guitar and several videos due to target person invisibility. All videos are up-sampled to 332 frames for action recognition. In the proposed single-shot cross-view setting, videos from one single view in the training set are leveraged for training and videos from all views are used for testing. The final performance is measured by the average top-1 accuracy over all views.

NTU-RGB+D. This dataset [38] contains 56,000 video clips in 60 action classes performed by 40 actors captured in-lab environments. Each clip has at most two subjects. Three cameras are used for recording different horizontal views simultaneously, and each action is performed twice towards the left and right cameras, respectively. Thus, there are six views contained in this dataset. Following [50], we pad every clip by replaying the sequence from the start to have 300 frames. Furthermore, we only use single-person action categories and the main actor in each video.

There are two common evaluation benchmarks [38] for action recognition on this dataset. In cross-subject benchmark, 40 subjects are split into training and testing groups, where each group consists of 20 subjects. In cross-view benchmark, training clips come from the second and third cameras, while the evaluation clips are all from the first camera. In this work, we introduce a new evaluation setting for single-shot cross-view action recognition. Specifically, we divide the training set of cross-subject benchmark into six splits according to cameras and replication numbers so that actions performed from only one view are contained in each split, and testing groups including all views and remaining subjects are used for evaluation. We report the average performance of all models trained using the six splits.

4.2. Implementation Details

Our approach does not require a particular 2D pose estimator, as long as it is reasonable accurate. We use [32] in our experiments. All detected keypoints of a 2D pose are then converted into a skeleton representation that consists of 13 joints according to the keypoint definition in [43]. We treat two poses as a positive pair if they are projected from the same 3D pose.

Camera Augmentation. We perform camera augmentation to improve the model robustness to large variations in camera viewpoints when applied to downstream tasks. When we train only with detected 2D keypoints in training images, we are constrained to the camera views in the training set. To reduce overfitting to these camera views, we perform camera augmentation by generating projected 2D keypoints from 3D poses at random views. For random rotation in camera augmentation, we follow [43] and uniformly sam-

Methods	VD	Left	Right	Front	Back	Average
Res-TCN [18]		86.83 ± 0.50	90.80 ± 1.09	75.99 ± 1.22	75.23 ± 3.05	82.21 ± 0.71
Temporal ConvNet		82.78 ± 1.38	88.78 ± 1.35	72.69 ± 1.83	69.70 ± 1.52	78.49 ± 0.91
Auto-Encoder	✓	85.89 ± 0.46	90.55 ± 0.85	77.45 ± 2.02	87.98 ± 0.93	85.47 ± 0.75
VAE [19]	✓	87.22 ± 1.19	92.14 ± 0.39	75.87 ± 2.14	88.76 ± 1.14	86.00 ± 0.72
β-VAE [1, 14]	✓	85.86 ± 1.27	90.03 ± 1.24	75.83 ± 1.06	83.56 ± 1.58	83.82 ± 0.60
InfoNCE [30]	✓	87.47 ± 0.85	89.25 ± 0.74	73.30 ± 0.59	83.05 ± 1.55	83.27 ± 0.55
DIM [15]	✓	81.67 ± 0.70	82.64 ± 0.67	76.08 ± 1.75	80.17 ± 1.29	80.14 ± 0.67
Pr-VIPE [43]		90.06 ± 0.38	89.36 ± 0.68	85.11 ± 0.69	91.58 ± 0.76	89.03 ± 0.40
CV-MIM	✓	**91.82 ± 0.30**	**93.73 ± 0.27**	**88.81 ± 0.53**	**92.65 ± 0.32**	**91.75 ± 0.24**

Table 1. Classification accuracy (%) and standard deviation of models on Penn Action [53] with the setting of single-shot cross-view action recognition. Each time, models are trained using one of the left, right, front, and back views, and evaluated on all four views. We highlight view-disentangled (VD) methods. Results are averaged over five runs; best performances are highlighted in bold.

Methods	VD	RGB	Flow	Pose	Accuracy
Nie *et al.* [28]		✓		✓	85.5
Cao *et al.* [5]			✓	✓	95.3
		✓	✓		98.1
Du *et al.* [9]		✓	✓	✓	97.4
Liu *et al.* [24]		✓		✓	91.4
Luvizon *et al.* [26]		✓		✓	98.7
Res-TCN [18]				✓	98.8
Temporal ConvNet				✓	98.5
Auto-Encoder	✓			✓	97.7
VAE [19]	✓			✓	97.6
β-VAE [1, 14]	✓			✓	97.7
InfoNCE [30]	✓			✓	97.5
DIM [15]	✓			✓	97.3
Pr-VIPE [43]				✓	**98.4**
CV-MIM	✓			✓	<u>98.1</u>

Table 2. Comparisons of top-1 action recognition accuracy (%) on Penn Action [53]. The check marks indicate the input to each method, including image pixels (RGB), optical flow (Flow), and model-estimated 2D pose (Pose). Top two performances of representation learning models are highlighted in bold and underline.

ple azimuth angle between $\pm 180°$, elevation between $\pm 30°$, and roll between $\pm 30°$. During model training, we use an even mixture of detected and projected 2D keypoints from different views to form positive 2D pose pairs.

Network Training. The backbone network architecture for our model is based on [27]. We use $d = 32$ for both pose and view representations as a good trade-off between model size and accuracy. The discriminator function f in Eq. (6) is implemented by the encode-and-dot-product architecture [15] which enables us to use large numbers of positive/negative samples, and mixture-of-experts [40] is employed for representation fusion. To weigh different losses in Eq. (11), we set $\lambda_1 = 5.0$, $\lambda_2 = 0.5$, and $\lambda_3 = 1.0$ such that all the loss terms have the same order of magnitude and did not densely tune them. Our implementation is in TensorFlow, and the model is trained with Tesla V100

GPUs. AdaGrad [10] with a learning rate of 0.02 is used for optimization, and we train the model for 5×10^6 iterations with mini-batches of size 256. The encoder operates on a single pose and is fixed for downstream tasks. More details on network architectures and model training are provided in the supplementary materials.

4.3. Action Recognition

We evaluate our approach in the downstream task of action recognition over a variety of settings including full-supervision, the proposed single-shot cross-view scenario, and limited-supervision.

Baselines. For fully-supervised baselines taking 2D poses, we compare our approach with three state-of-the-art methods based on temporal convolutions: Res-TCN [18], ST-GCN [50], and HCN [20]. We also compare to other state-of-the-art methods using different input modalities.

For representation learning, generative models are commonly used for building representations. Although their target domains are different, they usually optimize the following objective based on cross reconstruction [29, 33, 35, 36] for view-disentangled representation learning:

$$\min_{E,G} \left[\|\boldsymbol{x} - G(\boldsymbol{z}_p, \boldsymbol{z}_v)\|_2^2 + \|\boldsymbol{x} - G(\boldsymbol{z}_p^+, \boldsymbol{z}_v)\|_2^2 \right], \quad (12)$$

where G denotes the decoder, and $\| \cdot \|_2$ denotes the $L2$ distance. In the same spirit, we implement three cross-reconstruction baselines according to Eq. (12) using auto-encoder, VAE [19] and β-VAE [1, 14]. Their backbone networks of the encoder and decoder are both based on [27]. Moreover, we implement two MI-based counterparts optimizing Eq. (2) through InfoNCE [30] and DIM [15]. All these representation learning approaches predict both view and pose representations like our algorithm and thus are our main competing approaches. We also include Pr-VIPE [43] for comparison since they learn view-invariant embeddings for human poses as well, but we note that they do not produce view representations.

Methods	VD	C1-R1	C1-R2	C2-R1	C2-R2	C3-R1	C3-R2	Average
Res-TCN [18]		40.6 / 69.6	39.9 / 66.8	30.7 / 53.3	48.1 / 74.5	48.2 / 75.5	29.8 / 55.3	39.6 / 65.8
ST-GCN [50]		43.3 / 73.1	44.1 / 72.8	30.7 / 57.5	51.4 / 79.7	53.1 / 82.5	29.7 / 59.2	42.1 / 70.8
HCN [20]		52.5 / 80.3	49.8 / 76.9	37.2 / 63.8	55.5 / 86.7	55.3 / 83.8	39.4 / 69.3	48.3 / 76.8
Auto-Encoder	✓	43.6 / 75.6	39.6 / 74.9	29.4 / 61.7	45.7 / 77.7	41.3 / 73.0	33.4 / 70.0	38.8 / 72.2
VAE [19]	✓	50.2 / 81.5	50.4 / 80.4	38.5 / 70.4	54.1 / 82.8	54.7 / 82.1	37.6 / 69.6	47.6 / 77.8
β-VAE [1, 14]	✓	49.1 / 80.1	49.4 / 80.7	41.0 / 72.9	52.2 / 82.0	52.7 / 81.4	35.8 / 70.3	46.7 / 77.9
InfoNCE [30]	✓	43.0 / 75.7	43.0 / 74.3	36.4 / 65.9	46.4 / 76.7	49.0 / 77.4	36.3 / 68.2	42.3 / 73.0
DIM [15]	✓	42.1 / 74.4	42.5 / 74.3	32.3 / 61.8	45.7 / 74.0	43.9 / 71.3	33.3 / 65.6	40.0 / 70.2
Pr-VIPE [43]		56.1 / 85.8	57.9 / 85.5	50.7 / 84.1	57.3 / 84.5	55.9 / 84.1	50.9 / 83.5	54.8 / 84.6
CV-MIM	✓	**58.9 / 87.4**	**59.9 / 87.3**	**52.6 / 84.8**	**58.3 / 84.9**	**57.9 / 85.0**	**51.9 / 84.9**	**56.6 / 85.7**

Table 3. Results of top-1 and top-5 action recognition accuracy (%) on NTU-RGB+D [38] with the setting of single-shot cross-view action recognition. C1, C2, and C3 are the camera identifiers; R1 and R2 are the replication numbers; one combination of them forms a unique camera view. Each time, models are trained using one view, and evaluated on all six views. We highlight view-disentangled (VD) representation learning methods. Best performances are highlighted in bold.

Methods	VD	RGB	Depth	Pose	CS	CV
Vyas et al. [45]		✓			83.3	89.3
			✓		70.8	77.5
Res-TCN [18]				✓	82.2	90.2
ST-GCN [50]				✓	82.0	91.3
HCN [20]				✓	81.0	90.1
Auto-Encoder	✓			✓	62.5	71.8
VAE [19]	✓			✓	76.8	88.7
β-VAE [1, 14]	✓			✓	76.1	88.8
InfoNCE [30]	✓			✓	74.1	82.7
DIM [15]	✓			✓	73.6	82.4
Pr-VIPE [43]				✓	<u>77.7</u>	**89.7**
CV-MIM	✓			✓	**77.8**	<u>89.5</u>

Table 4. Comparisons of top-1 action recognition accuracy (%) on NTU-RGB+D [38] with Cross-Subject (CS) and Cross-View (CV) settings. The check marks indicate the input to each method, including image pixels (RGB), depth, and model-estimated 2D pose (Pose). We highlight view-disentangled (VD) representation learning methods. Top two performances of representation learning models are highlighted in bold and underline, respectively.

Figure 3. Recognition accuracy when limited supervisions are provided on Penn Action [53] (**top**) and NTU-RGB+D [38] (**bottom**).

Results on Penn Action. We start by evaluating our method on Penn Action [53]. A simple temporal convolution network is used for aggregating temporal features from per-frame pose representations. See supplementary materials for detailed architecture and training setup. Tables 1 and 2 show the results in the single-shot cross-view and fully-supervised settings, respectively.

We observe that our approach outperforms other view-disentangled as well as fully-supervised methods by a large margin in the single-shot setting and presents the lowest variance in accuracy. It is also worth mentioning that our results are substantially better than those baselines directly optimizing Eq. (2), which demonstrates the effectiveness of our refined formulation proposed in Eq. (11). Additionally, we match the state of the art in the fully-supervised setting
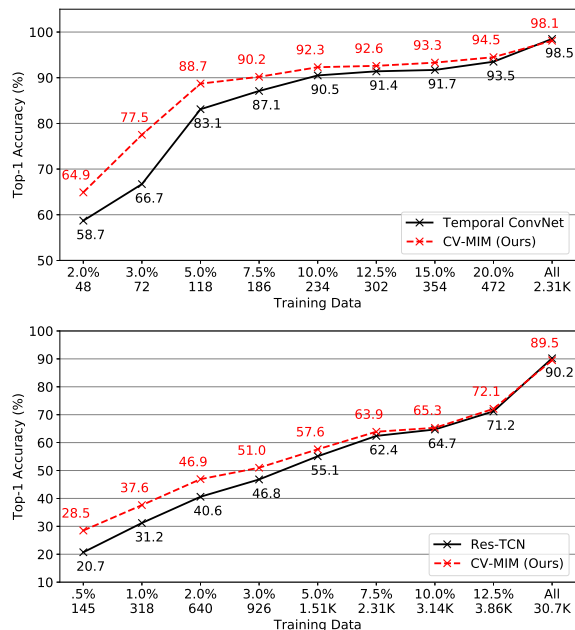
and even yield better results than models using multiple input modalities.

Results on NTU-RGB+D. We continue to experiment on NTU-RGB+D [38]. The results under the same settings are reported in Tables 3 and 4, respectively. We observe that our model achieves the best performance in the single-shot setting while achieving competitive results in the fully-supervised setting. Interestingly, some cross-reconstruction models fail because of the large variances in poses and viewpoints present in this dataset. In contrast, our model is robust to these changes. From Table 4, we also observe that there is in general a considerable performance gap between the fully-supervised methods and our

Figure 4. Nearest neighbors in the representation space using subjects S9 and S11 on Human3.6M [17]. The first row uses pose representations; the second uses view representations. On each row, we show the query on the left and its 5 nearest neighbors on the right.

Methods	Concat	Product	Mixture	Accuracy
CV-MIM (full)	✓			90.5
		✓		90.2
			✓	**91.8**
w/o $\mathcal{L}_{inter}$			✓	86.1
w/o $\mathcal{L}_{prior}$			✓	89.3

Table 5. Ablation study on the fusion operation and two regularization losses used in Eq. (11) on Penn Action [53].

method in the cross-subject experiment but not in the cross-view experiment. This indicates our learned representation is not subject-invariant, which could be potentially solved by training with more subjects or augmented skeletons.

Training with Limited Supervisions. We further investigate model performances when supervised data is limited. In this experiment, we use the same setup as the fully-supervised setting described above except that the amount of supervised training samples is varied. We report the results in Fig. 3. We find that our model consistently improves the fully-supervised baselines by a notable margin when only a limited number of training samples are available. This shows that the learned pose representations can capture the semantics of 2D poses in a meaningful way which reduces the amount of supervision needed for the downstream task. We also provide additional comparisons with other representation learning methods under the same setting in the supplementary materials.

4.4. Ablation Study

We perform the ablative analysis to better understand the design choices of our approach from two perspectives: the fusion operation to combine view and pose representations and the effectiveness of two regularization losses used in Eq. (11). We explore three choices of the fusion operation

in this experiment: concatenation, product-of-experts [49], and mixture-of-experts [40]. Table 5 shows the results on Penn Action [53] in the single-shot setting. We can see that our model is robust to the form of fusion operation thanks to the disentanglement loss. We select mixture-of-experts as the default fusion operation due to its highest performance. Furthermore, removing either regularization loss in Eq. (11) results in a significant performance downgrade, which demonstrates their effectiveness.

4.5. Qualitative Results

Last but not least, we show qualitative results when using the learned representations of our model for nearest neighbor retrieval. In Fig. 4, we show that our pose representations can successfully find similar poses from different views in the testing set of Human3.6M [17]. Interestingly, we also show that the learned view representations can retrieve frames captured from a viewpoint that is very similar to the query while the poses are different. More visual results are provided in the supplementary materials.

5. Conclusion

We present CV-MIM, a representation learning approach to encode both pose-dependent and view-dependent representations for 2D human poses by maximizing cross-view MI. We further motivate two regularization losses to encourage disentanglement and smoothness of the learned representations from a theoretical perspective. We show that using the learned pose representation achieves significant improvement from existing representation learning methods on downstream action recognition tasks, and even outperforms fully-supervised baselines in many settings. We also demonstrate the learned view representation can be directly applied to similar view retrieval among different poses. CV-MIM focuses on single person representation, and for future work, we will investigate its multi-person extension.

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *Int. Conf. Learn. Represent.*, 2017.

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Adv. Neural Inform. Process. Syst.*, pages 15535–15545, 2019.

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML*, pages 531–540, 2018.

[4] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[5] Congqi Cao, Yifan Zhang, Chunjie Zhang, and Hanqing Lu. Body joint guided 3D deep convolutional descriptors for action recognition. *IEEE Transactions on Cybernetics*, 48(3):1095–1108, 2017.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[7] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. CLUB: A contrastive log-ratio upper bound of mutual information. In *ICML*, 2020.

[8] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[9] Wenbin Du, Yali Wang, and Yu Qiao. RPAN: An end-to-end recurrent pose-attention network for action recognition in videos. In *Int. Conf. Comput. Vis.*, pages 3725–3734, 2017.

[10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

[11] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1801–1810, 2019.

[12] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019.

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9729–9738, 2020.

[14] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Int. Conf. Learn. Represent.*, 2017.

[15] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *Int. Conf. Learn. Represent.*, 2019.

[16] Chih-Hui Ho, Pedro Morgado, Amir Persekian, and Nuno Vasconcelos. PIEs: Pose Invariant Embeddings. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12377–12386, 2019.

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2013.

[18] Tae Soo Kim and Austin Reiter. Interpretable 3D human action analysis with temporal convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 1623–1631, 2017.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Int. Conf. Learn. Represent.*, 2014.

[20] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *IJCAI*, pages 786–792, 2018.

[21] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Int. Conf. Comput. Vis.*, pages 1446–1455, 2019.

[22] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Unsupervised learning of view-invariant action representations. In *Adv. Neural Inform. Process. Syst.*, pages 1254–1264, 2018.

[23] Jingen Liu, Mubarak Shah, Benjamin Kuipers, and Silvio Savarese. Cross-view action recognition via view knowledge transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3209–3216, 2011.

[24] Mengyuan Liu and Junsong Yuan. Recognizing human actions as the evolution of pose estimation maps. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1159–1168, 2018.

[25] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10955–10964, 2019.

[26] Diogo Luvizon, David Picard, and Hedi Tabia. Multi-task deep learning for real-time 3D human pose estimation and action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[27] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *Int. Conf. Comput. Vis.*, pages 2640–2649, 2017.

[28] Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu. Joint action recognition and pose estimation from video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1293–1301, 2015.

[29] Qiang Nie, Ziwei Liu, and Yunhui Liu. Unsupervised 3D human pose representation with viewpoint and pose disentanglement. In *Eur. Conf. Comput. Vis.*, 2020.

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[31] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. In *Adv. Neural Inform. Process. Syst.*, pages 15604–15614, 2019.

[32] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4903–4911, 2017.

[33] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *Int. Conf. Comput. Vis.*, pages 1623–1632, 2017.

[34] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(12):2430–2443, 2016.

[35] Helge Rhodin, Victor Constantin, Isinsu Katircioglu, Mathieu Salzmann, and Pascal Fua. Neural scene decomposition for multi-person motion capture. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7703–7713, 2019.

[36] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3D human pose estimation. In *Eur. Conf. Comput. Vis.*, pages 750–767, 2018.

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[38] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1010–1019, 2016.

[39] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang. Deep multimodal feature analysis for action recognition in RGB+D videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(5):1045–1058, 2017.

[40] Yuge Shi, N Siddharth, Brooks Paige, and Philip Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Adv. Neural Inform. Process. Syst.*, pages 15718–15729, 2019.

[41] Nicki Skafte and Søren Hauberg. Explicit disentanglement of appearance and perspective in generative models. In *Adv. Neural Inform. Process. Syst.*, pages 1018–1028, 2019.

[42] Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *Int. Conf. Learn. Represent.*, 2020.

[43] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *Eur. Conf. Comput. Vis.*, 2020.

[44] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Eur. Conf. Comput. Vis.*, 2020.

[45] Shruti Vyas, Yogesh S Rawat, and Mubarak Shah. Multi-view action recognition using cross-view video prediction. In *Eur. Conf. Comput. Vis.*, 2020.

[46] Dongang Wang, Wanli Ouyang, Wen Li, and Dong Xu. Dividing and aggregating network for multi-view action recognition. In *Eur. Conf. Comput. Vis.*, pages 451–467, 2018.

[47] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2649–2656, 2014.

[48] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pages 9929–9939, 2020.

[49] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Adv. Neural Inform. Process. Syst.*, pages 5575–5585, 2018.

[50] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pages 7444–7452, 2018.

[51] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(8):1963–1978, 2019.

[52] Pengfei Zhang, Jianru Xue, Cuiling Lan, Wenjun Zeng, Zhanning Gao, and Nanning Zheng. Adding attentiveness to the neurons in recurrent neural networks. In *Eur. Conf. Comput. Vis.*, pages 135–151, 2018.

[53] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Int. Conf. Comput. Vis.*, pages 2248–2255, 2013.

[54] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *Eur. Conf. Comput. Vis.*, pages 387–403, 2018.

[55] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3D human pose regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3425–3435, 2019.

[56] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *IEEE Trans. Image Process.*, 28:4500–4509, 2019.