DVANet: Disentangling View and Action Features for Multi-View Action Recognition

Nyle Siddiqui, Praveen Tirupattur, Mubarak Shah

Center for Research in Computer Vision, University of Central Florida
nyle.siddiqui@ucf.edu, praveen.tirupattur@ucf.edu, shah@crcv.ucf.edu

Abstract

In this work, we present a novel approach to multi-view action recognition where we guide learned action representations to be separated from view-relevant information in a video. When trying to classify action instances captured from multiple viewpoints, there is a higher degree of difficulty due to the difference in background, occlusion, and visibility of the captured action from different camera angles. To tackle the various problems introduced in multi-view action recognition, we propose a novel configuration of learnable transformer decoder queries, in conjunction with two supervised contrastive losses, to enforce the learning of action features that are robust to shifts in viewpoints. Our disentangled feature learning occurs in two stages: the transformer decoder uses separate queries to separately learn action and view information, which are then further disentangled using our two contrastive losses. We show that our model and method of training significantly outperforms all other uni-modal models on four multi-view action recognition datasets: NTU RGB+D, NTU RGB+D 120, PKU-MMD, and N-UCLA. Compared to previous RGB works, we see maximal improvements of 1.5%, 4.8%, 2.2%, and 4.8% on each dataset, respectively. Our code can be found here: *https://github.com/NyleSiddiqui/MultiView_Actions*

Introduction

Human action recognition (HAR) has played an integral role in the development and progression of computer vision research. HAR has established itself as a critical intersection between academic research and real-world implementation due to its practical application to a myriad of domains, such as human-computer interaction (Chakraborty et al. 2018), road safety (Ismail 2010), video understanding (Hussain et al. 2021), security/surveillance (Black, Ellis, and Rosin 2002), and more. Consequently, there are a plethora of large-scale datasets that have been collected over the years to provide ample resources for researchers to advance this historically active field. Furthermore, HAR datasets have grown in difficulty proportionally to the astounding capabilities of deep learning in computer vision, such as the introduction of additional modalities, increased number of subjects and actions, and multiple viewpoints.

In specific regards to multi-view action recognition, it is currently a vital and heavily researched variant of HAR as it is more challenging and representative of real-world scenarios. For example, HAR datasets are commonly collected in an in-the-wild manner, where labelled actions are collected from viewpoints that are not explicitly controlled. Since single-view action recognition is usually the main learning objective in standard HAR, it is not guaranteed that the learned action representations of such models are robust to changes in viewpoints. However, in the previously mentioned real-world scenarios, videos of an action being performed are synchronously captured from multiple known viewpoints. Thus, multi-view action recognition has emerged as a more modern, challenging, and realistic version of HAR (Liu et al. 2016; Gao et al. 2019). Additionally, many of these in-the-wild scenarios may lack the equipment or controlled collection that is required to capture additional modalities besides RGB frames, such as skeletal or depth (Sun et al. 2022). This proves to be a significant obstacle for current state-of-the-art multi-modal methods, as many of these solutions are heavily reliant on these modalities.

There are also many tangential areas of deep learning that affect the current approaches of multi-view action recognition. For example, a major obstacle in supervised deep learning is ensuring that learned features do not contain any irrelevant or confounding information with respect to the learning objective. This can be difficult to achieve since we cannot fully control how supervised models learn and represent information in their feature space. This can lead to learned features becoming 'entangled' with irrelevant information, thereby degrading performance. An example of this phenomenon in multi-view action recognition can be seen in Figure 1, where view and action information can potentially become entangled in the feature space. However, large strides have been made in representation learning to close these gaps, including disentangled representation learning (DRL) (Liu et al. 2018; Higgins et al. 2018). DRL aims to improve the learned features for a training objective by enforcing strict separation between relevant and irrelevant information, improving overall performance.

Therefore, this work proposes a novel approach specifically towards multi-view action recognition in which we train a transformer decoder-based model for view-invariant action recognition using only a single modality. Our train-
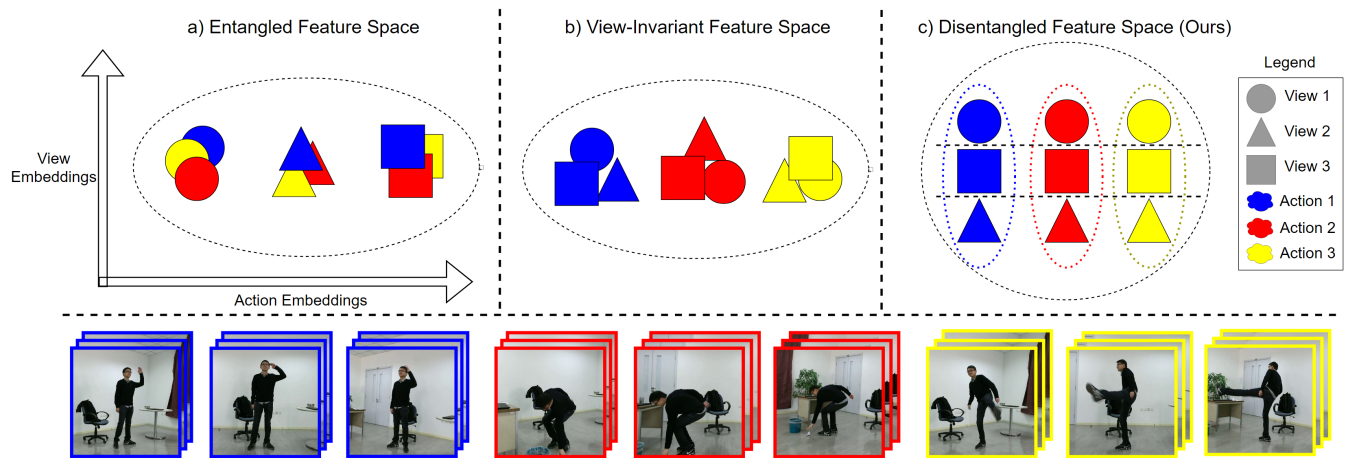
Figure 1: Conceptual visualization of learned action and view features plotted on the x- and y-axes, respectively, for different learning objectives. a) Traditional feature learning for multi-view action recognition may lead to view features becoming entangled in the action embedding space, causing different action classes to erroneously cluster due to similar viewpoints. b) View-invariant action features will properly cluster by action class, but the disentangled view features do not cluster properly. c) Our method disentangles the view features from the action features while still retaining the structure of both embedding spaces, improving performance on unseen viewpoints (see Fig. 3).

ing method leverages disentangled representation learning by guiding our model to separate the action-relevant features from the view-relevant features. This enables our model to learn better quality action features that are robust to changes in viewpoints. We implement this disentanglement in two stages. In the first stage, we introduce the use of a singular view query in our transformer decoder to capture and disentangle view information from the learned action features. In the second stage, we instill relative comparisons between features using two supervised contrastive losses to assist in learning disentangled features. Standard cross-entropy learning only challenges a model to produce features that map a singular sample to the correct label. The model may implicitly learn the relative inter- and intra-class similarities and differences in the dataset, however this is not explicitly enforced and thus not guaranteed. On the other hand, supervised contrastive learning utilizes additional samples during training to guide a model to learn relatively identical features for similar samples (positives) and relatively disjoint features for dissimilar samples (negatives). We show that our novel method of training, paired with our novel architecture, mitigates the degradation in multi-view action recognition performance stemming from the noise imparted by multiple viewpoints. Our main contributions are as follows:

- A novel configuration of transformer decoder queries that enforce disentanglement of learned action and view features for view-invariant action recognition.

- Two supervised contrastive losses and a query orthogonality loss to further supplement our disentangled representation learning.

- Unlike most methods, our approach is compatible with both the RGB or skeleton modality and is shown to outperform all unimodal state-of-the-art models on four different multi-view action recognition datasets.

Related Work

Multi-View Action Recognition

Many recent action recognition works focus on tackling the problem of multi-view action recognition due to the modern and realistic challenge it poses in conjunction with its relevance in representation learning (Kong et al. 2017; Bian et al. 2023). The current dominant approaches consist of skeletal, RGB, or multi-modal methods; however, they are not without their limitations. Skeletal-based methods (Chen et al. 2021; Shi et al. 2021; Song et al. 2017) are the most common approach to multi-view action recognition due to their quality representation of motion, removal of irrelevant information like background and clothing, and widespread availability of accurate skeleton ground truth labels in large-scale datasets (Liu et al. 2017; Xia, Chen, and Aggarwal 2012; Liu et al. 2019). Additionally, the compatibility of graph convolutional methods with skeleton-based action recognition led to widespread use in the literature (Chi et al. 2022; Song et al. 2022; Cheng et al. 2020). RGB-based action recognition is more sparse in the literature due to its lack of 3D structure, usually only being used in addition to other modalities (Wang et al. 2019; Cheng et al. 2022; Das et al. 2020). Contrary to these previous works, we demonstrate how our RGB-based model is able to achieve state-of-the-art multi-view action recognition performance, even over skeletal-based models.

In addition to uni-modal approaches, multi-modal learning has shown promising results due to the additional quantity and modes of information provided during training. (Bruce et al. 2022) proposed a multi-modal approach to action recognition, named MMNet, in which they fuse skeletal and RGB-based features in a complementary manner to learn better action representations. MMNet does show impressive results on many multi-modal action recognition

datasets, however we exhibit that we are able to achieve competitive results with only the RGB modality. Furthermore, our approach does not require the additional extraction of skeletal features, allowing for a more lightweight approach without any significant sacrifice in performance; an attribute not commonly seen with other RGB-based methods (Wang et al. 2018b; Lin et al. 2020; Shah et al. 2023). Previous RGB methods have similarly explored contrastive learning for multi-view action recognition, but still underperform when compared to skeletal-based models, whereas our approach outperforms all uni-modal methods and sets multiple new state of arts.

Learning view-invariant representations for multi-view action recognition has also been explored before for both uni- and multi-modal approaches (Li et al. 2018b; Das and Ryoo 2023; Ji et al. 2021; Bian et al. 2023), and by using either solely convolutional or transformer-based architectures (Cheng et al. 2022; Vyas, Rawat, and Shah 2020; Ji et al. 2021). Our approach uses a hybrid architecture to limit each of these respective shortcomings, while also novelly introducing a specific configuration of queries for transformer decoders that facilitates feature disentanglement. We show that these novelties significantly contribute to our state-of-the-art results and strong improvements over previous works.

Disentangled Representation Learning

Disentanglement has been applied in multi-view action recognition at both the representation learning level (Zhao et al. 2021; Guo et al. 2022) and higher (Liu et al. 2020), as well as other domains (Zhang et al. 2019; Jin et al. 2022; Zhou et al. 2022b,a). Similar to our method, (Tran, Yin, and Liu 2017) leveraged DRL for pose-invariance, but for face recognition. They also approach DRL from a joint generative-discriminative approach, which has been seen previously (Zheng et al. 2019), in addition to purely generative methods (Karras, Laine, and Aila 2019). While our approach is purely discriminative, DRL's vast applicability across many domains is a testament to its capability to improve learned representations for specific tasks. However, improving learned representations is not limited to DRL. (Guo et al. 2022) explored improving skeleton-based action recognition representations using a self-supervised contrastive learning method. (Wang et al. 2018a), somewhat similar to our method, uses view classification as an auxiliary task to assist in learning view-invariant representations.

Methodology

Motivation

It is clear that the strongest factor of difficulty in multi-view action recognition, compared to general HAR, is the multiple viewpoints from which an action is captured. Supervised deep learning models trained on single-view action recognition datasets are not incentivized to learn as robust action representations as in the multi-view setting, since the performed actions are viewed from a static and homogeneous viewpoint. With the introduction of multiple viewpoints, the visibility and information of a performed action is variable, and thus requires more generalized and robust features to account for these perturbations. In order to further advance general video understanding, it is imperative to develop supervised techniques that can process information from multiple viewpoints since even state-of-the-art single-view models will still fail in this facet.

Our Approach

To this end, we propose DVANet to overcome these challenges, the full details of which can be seen in Figure 2. We use a hybrid transformer decoder architecture with a 3D-CNN encoder to instill both local and global reasoning into our model. For the first stage of disentanglement, our decoder consists of multiple action queries to capture all action-relevant information, and a singular view query to capture all view-relevant information. In addition to standard cross-entropy loss, we utilize two supervised contrastive losses as a second stage of disentanglement to assist in the learning of view-invariant action features. These losses help drive feature disentanglement by enforcing the learned action and view features to only contain action and view information, respectively, and minimize leakage. We expect the action queries in the decoder to encapsulate all action-relevant information in a video without any perturbations due to changes in the camera view. Moreover, many previous methods rely on skeletal data in addition to RGB data to support the learning of view-invariant action features, whereas we achieve improved performance on the same task using only RGB data. We show that our novel architecture paired with our simple training method leads to significantly improved results over both RGB-based models and skeleton-based models alike on four different multi-view action recognition datasets: NTU RGB+D, NTU RGB+D 120, PKU-MMD, and N-UCLA.

Input Videos During training, we pass a triplet of videos through our network. There is an anchor video, $\mathbf{a}$, consisting of an action being performed from some known labeled viewpoint. We then retrieve a video of a different action captured from the same viewpoint as $\mathbf{a}$ and denote this video as $\mathbf{sv}$ (same view). Lastly, we retrieve a video of the same action as $\mathbf{a}$ captured from a different viewpoint, denoted as $\mathbf{sa}$ (same action). We construct the triplets in this manner to provide positive and negative samples for our supervised contrastive feature learning, discussed later in this section. Providing additional positive and negative samples enables the model to gain a more holistic understanding of the learning objective through relative comparisons of samples, as opposed to standard single-label training.

Feature Extraction Our model consists of a backbone encoder paired with a transformer decoder. The backbone can be convolutional- or transformer-based, but an unfrozen, pretrained R3D backbone yielded best results. We firstly extract features using our backbone, which takes as input a video clip $\mathbf{v} \in \mathbb{R}^{(T \times C \times H \times W)}$, where T is the number of frames, and $C, H,$ and W are the number of channels, height, and width of each frame respectively. The backbone encoder produces a feature map $\mathbf{f} = ENC(\mathbf{v}) \in \mathbb{R}^{(T_s \times D)}$ as output, where D is the size of the hidden dimension and T_s is the number of features for the entire video clip after
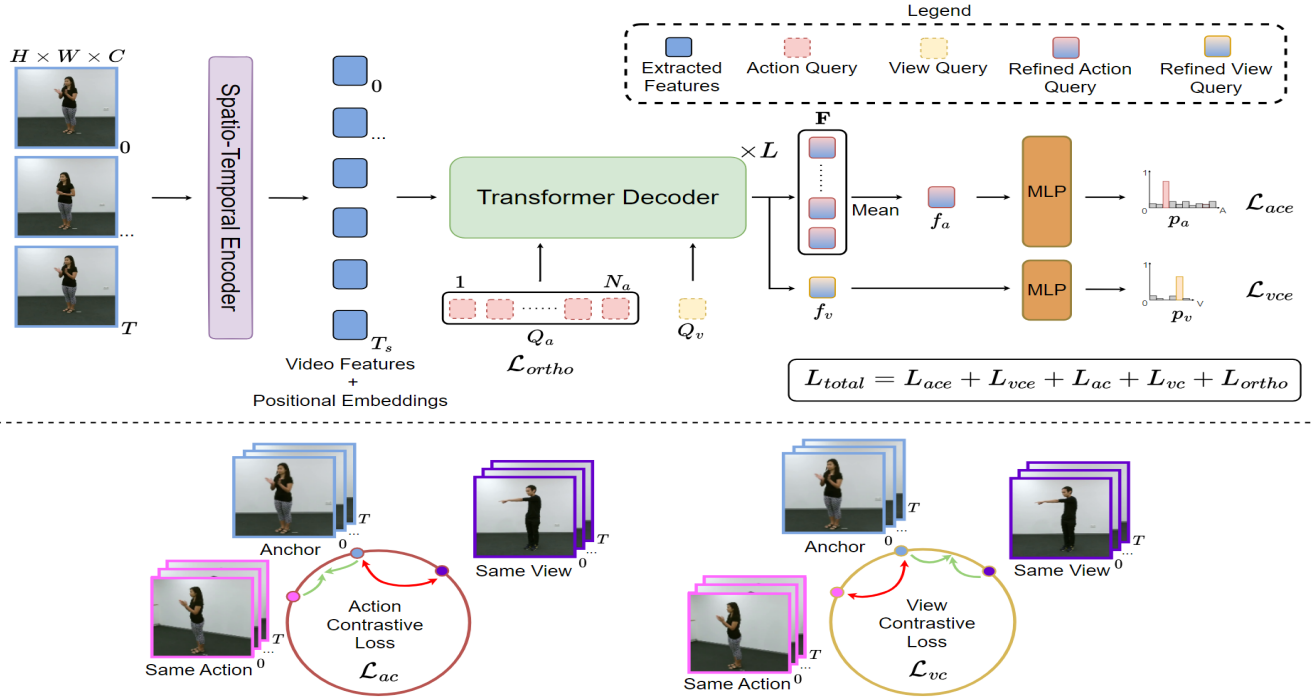
Figure 2: First, our spatio-temporal encoder extracts global features from a video. A transformer decoder is used to separately extract action and view features with the help of learnable action ($\mathbf{Q}_a$) and view ($\mathbf{Q}_v$) queries. In addition to the classification loss to predict the action and view, we also introduce contrastive loss to learn disentangled view and action representations.

temporal down-sampling. A learnable positional encoding is appended to this output, as is required, before it is passed to the transformer decoder for feature disentanglement.

Feature Disentanglement Our decoder takes the extracted spatio-temporal features from the backbone encoder as input and performs the first stage of disentanglement. The decoder consists of N_a learnable action queries and a singular view query. Through our loss formulation, the action queries $\mathbf{Q}_a \in \mathbb{R}^{(N_a \times D)}$ learn to encapsulate the semantic information in a video that is pertinent only to the action being performed, irrespective of the view it was captured from. Contrastively, the singular view query $\mathbf{Q}_v \in \mathbb{R}^{(1 \times D)}$ is expected to only learn view-relevant information in a video, irrespective of the action being performed. This view query is meant to capture any perturbations that would otherwise occur in the action embedding space due to changes in the camera viewpoint, thereby 'disentangling' this noise from the learned action features. In addition to the contrastive learning, we use view classification as a training objective for this view query to learn view-relevant information. Thus, we are encouraging the action queries to produce view-invariant action features, since the view query captures any view-relevant information.

The extracted features from our backbone encoder are then used as input for the transformer decoder, where these learnable queries are utilized to produce the decoder's output of refined features:

$$\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_{N_a}] \in \mathbb{R}^{(N_a \times D)}, \ \mathbf{f}_v \in \mathbb{R}^{(1 \times D)}$$

where $[\mathbf{f}_1, \cdots, \mathbf{f}_{N_a}]$ are the refined action features and $\mathbf{f}_v$ is the final learned view feature. We then take the average across the refined action features $\mathbf{F}$ to impart all learned query information into a final action feature, $\mathbf{f}_a$, where $\mathbf{f}_a \in \mathbb{R}^{(1 \times D)}$. Lastly, $\mathbf{f}_a, \mathbf{f}_v$ are passed to separate fully-connected linear layers L_a, L_v that map the action and view features to their corresponding prediction dimensions, respectively. Action and view logits, $\mathbf{p}_a, \mathbf{p}_v$ are attained by

$$\mathbf{p}_a = L_a(\mathbf{f_a}) \in \mathbb{R}^A \text{ and } \mathbf{p}_v = L_v(\mathbf{f_v}) \in \mathbb{R}^V,$$

where A and V are the number of actions and views.
Both the action and view branches of our model are tuned during training, but only the action branch is used during inference to produce action predictions. However, we cannot ensure that this architecture alone will result in disentangled action and view features without proper supervision. Therefore, we propose a combination of losses to properly enforce feature disentanglement.

Loss Formulation We incorporate standard cross entropy loss to train the features and linear layers using the corresponding action ($\mathcal{L}_{ace}$) or view ($\mathcal{L}_{vce}$) labels:

$$\mathcal{L}_{ace} = \sum_{n=1}^{N} \sum_{m=1}^{M} \log \mathcal{P}_m(\mathbf{p}_a), \qquad (1)$$

$$\mathcal{L}_{vce} = \sum_{n=1}^{N} \sum_{m=1}^{M} \log \mathcal{P}_m(\mathbf{p}_v), \qquad (2)$$

where N is the number of training samples, M is the number of actions (or the number of views in L_v), and $\mathcal{P}_m(\mathbf{p}_*)$ is the probability of each predicted class from the logits with ground truth class m.

Contrastive Feature Loss Using the aforementioned triplet construction, we propose two opposing supervised contrastive learning objectives. Both of these losses use the same triplet construction, however the difference resides in the designation of positive and negative samples. Firstly, the action contrastive loss (Eq. 3) prioritizes learning view-invariant action features by treating the **sa** video as a positive sample, and the **sv** video as the negative sample. Thus, the distance between outputted features from the anchor video **a**, and a video of the same action from a different view **sa**, are minimized. Conversely, the distance between outputted features from **a** and a video of a different action from the same viewpoint **sv** is maximized:

$$\mathcal{L}_{ac} = \sum_{n=1}^{N} \left[\delta + \mathcal{D}(\mathbf{f}_a, \mathbf{f}_{sa}) - \mathcal{D}(\mathbf{f}_a, \mathbf{f}_{sv}) \right], \qquad (3)$$

where $\mathbf{f}_a$, $\mathbf{f}_{sa}$, and $\mathbf{f}_{sv}$ denote the action feature outputs from the **a**, **sa**, and **sv** videos, respectively, $\mathcal{D}(\cdot, \cdot)$ is a distance function, and δ is a margin parameter. The learned action features are then expected to become view-invariant, since they are encouraged to be identical when resulting from two different videos that consist of the same action, regardless of which viewpoint it was recorded from. We reverse the positive and negative samples for the view contrastive loss:

$$\mathcal{L}_{vc} = \sum_{n=1}^{N} \left[\delta + \mathcal{D}(\mathbf{f}_a, \mathbf{f}_{sv}) - \mathcal{D}(\mathbf{f}_a, \mathbf{f}_{sa}) \right], \qquad (4)$$

such that the singular view query in the decoder learns to produce similar features for videos that are from the same viewpoint, regardless of the action being performed. Since our main objective is to improve multi-view action recognition, training our model to disentangle the action-relevant features from the view-relevant features in this manner provides cleaner, more accurate action representations that are robust to changes in the camera viewpoint.

Orthogonal Query Loss As seen in previous works (Fei et al. 2022; Zhang et al. 2021), initializing and enforcing orthogonality between queries in the decoder improves both model performance and numerical stability. Leveraging orthogonal queries aligns with our learning objective as we desire each action query to learn distinct, non-overlapping semantic information in the video to avoid redundancy. We use the cosine similarity between action queries $(\mathbf{q}_m, \mathbf{q}_n \in Q_a)$ to compute the loss and discourage any overlap between action queries during training:

$$\mathcal{L}_{ortho} = \sum_{m=0}^{N_a} \sum_{n \neq m} \left| \frac{\langle \mathbf{q}_m, \mathbf{q}_n \rangle}{||\mathbf{q}_m|| \, ||\mathbf{q}_n||} \right|. \qquad (5)$$

Our overall loss is then calculated as follows:

$$\mathcal{L} = \mathcal{L}_{ace} + \mathcal{L}_{vce} + \mathcal{L}_{ac} + \mathcal{L}_{vc} + \mathcal{L}_{ortho}.$$

Experimental Evaluation

Results

We show results on three large-scale, multi-view action recognition datasets (NTU RGB+D, NTU RGB+D 120, PKU-MMD) and one smaller scale dataset (N-UCLA) to exhibit the effectiveness of our approach with varying amounts of data. We show results on the standard cross-subject and cross-view evaluation protocols provided for each dataset. Additional analyses and details are in the supplement.

NTU RGB+D & NTU RGB+D 120 In Table 1, we provide a comparison of action test accuracy on NTU RGB+D with previous state-of-the-art (SOTA) unimodal models. In both the cross-subject and cross-view evaluation protocol, we achieve the highest performance and beat the previous best (which used the skeleton modality) in each protocol by 0.2% and 0.7%, respectively. When only compared to RGB models, we beat the previous best by an impressive 1.5% margin in cross-subject, and 0.2% in cross-view. An important distinction of our method is the use of view classification to better learn and identify view-relevant information, thereby making it easier to disentangle from our model's learned action features (see Table 5).

In Table 2, we observe similar superior performance of our method when applied to the larger dataset of NTU RGB+D 120. Our model achieves SOTA results compared to other unimodal models, achieving a 0.5% and 0.4% improvement over the SOTA skeleton-based model in the cross-subject and cross-setup protocols, respectively. More importantly, we significantly outperform RGB based models with large improvements of 4.8% and 4.1%, respectively. Smaller improvements are seen over skeleton-based methods due to the difference in quality of information between modalities, whereas our approach yields significantly large improvements over other RGB methods.

PKU-MMD Table 3 shows the comparisons of our results with previous works on PKU-MMD. Again we report large margins between our model and previous SOTA methods, with improvements of 2.2% and 1% on cross-subject and cross-view, respectively. The listed RGB-based methods even use additional depth information and still do not reach the our level of performance. Our model also beats all baselines in Table 3 when we train using the skeleton modality, which can be found in the supplement.

N-UCLA Compared to the other three large-scale datasets, N-UCLA is considerably smaller in both the number of videos and actions. Some works (Shah et al. 2023) evaluate their model using transfer learning and fine-tuning on this dataset, while others train fully from scratch. Previous works such as (Vyas, Rawat, and Shah 2020) are only able to obtain competitive results when using transfer learning, as training from scratch resulted in severely degraded performance. We train our model from scratch on this dataset and beat recent RGB works by a significant margin of 2.3% and 4.8%, as seen in Table 4. Our results indicate that our approach does not require mass amounts of data to achieve SOTA performance.

Model	CS	CV
(Liu et al. 2020) †	91.5	96.2
(Chen et al. 2021) †	92.4	96.8
(Chi et al. 2022) †	93.0	97.1
(Trivedi and Ravi 2023) †	92.9	96.7
(Duan et al. 2022) †	92.6	97.4
(Song et al. 2022) †	92.1	96.1
(Zhang, Lin, and Liu 2023) †	90.4	95.7
(Wang et al. 2018a) ‡	63.3	70.6
(Baradel et al. 2018) ‡	86.6	93.2
(Vyas, Rawat, and Shah 2020) ‡	82.3	86.3
(Piergiovanni and Ryoo 2021) ‡	-	93.7
(Cheng et al. 2022) ‡	91.9	95.4
(Das and Ryoo 2023) ‡	89.7	94.1
(Shah et al. 2023) ‡	91.4	98.0
DVANet ‡	**93.4** +1.5	**98.2** +0.2

Table 1: Comparison between DVANet and previous SOTA unimodal models on NTU RGB+D. Highest accuraices are in bold, and margin of improvement over RGB methods (‡) are underlined. † denotes skeleton-based methods.

Model	CS	XSet
(Liu et al. 2020) †	86.9	88.4
(Chen et al. 2021) †	88.9	90.6
(Chi et al. 2022) †	89.8	91.2
(Chi et al. 2022) †	88.5	89.7
(Song et al. 2022) †	88.7	88.9
(Duan et al. 2022) †	88.6	90.8
(Trivedi and Ravi 2023) †	89.4	90.6
(Zhang, Lin, and Liu 2023) †	85.6	87.5
(Das and Ryoo 2023) ‡	84.5	86.2
(Shah et al. 2023) ‡	85.6	87.5
DVANet ‡	**90.4** +4.8	**91.6** +4.1

Table 2: Comparison between DVANet and previous SOTA unimodal models on NTU RGB+D 120. Highest accuraices are in bold, and margin of improvement over RGB methods (‡) are underlined. † denotes skeleton-based methods.

Model	CS	CV
(Song et al. 2017) †	86.9	92.6
(Li et al. 2018a) †	92.6	94.2
(Elias, Jan, and Zezula 2019) †	86.5	92.2
(Wang et al. 2018b) ‡	85.0	85.7
(Lin et al. 2020) ‡	91.7	92.6
(Cheng et al. 2022) (TSN) ‡	92.6	92.1
(Cheng et al. 2022) (TSM) ‡	93.6	94.2
DVANet ‡	**95.8** +2.2	**95.2** +1.0

Table 3: Comparison between DVANet and previous SOTA unimodal models on PKU-MMD. Highest accuraices are in bold, and margin of improvement over RGB methods (‡) are underlined. † denotes skeleton-based methods.

Model	CS	CV
(Wang et al. 2018a) ‡	92.1	86.5
(Baradel et al. 2018) ‡	-	87.6
(Vyas, Rawat, and Shah 2020) ‡	87.5	83.1
(Das and Ryoo 2023) ‡	-	89.1
(Shah et al. 2023) ‡	-	91.7
DVANet ‡	**94.4** +2.3	**96.5** +4.8

Table 4: Comparison between DVANet and SOTA unimodal models on N-UCLA. Highest accuraices are in bold with margin of improvement over RGB methods (‡) underlined.

Discussion & Analysis

We provide multiple ablations and analyses on our losses and results. All ablations are performed using the cross-subject evaluation protocol on PKU-MMD. We use the cross-subject protocol since our model is trained in a supervised manner for view-invariant action recognition from seen viewpoints. However, Fig. 4 exhibits our model ability to implicitly perform on unseen viewpoints.

Ablations

Effectiveness of Using View Information When we remove the cross-entropy and contrastive losses related to view information, we observe a significant 1.8% drop in performance in the first row of Table 5. These two losses aid in the disentanglement of view information from the action features, and thus removing them expectedly degrades the quality of learned action features. This shows that multiple viewpoints does in fact perturb learned action features enough to drop performance.

Effectiveness of Contrastive Losses Rows 2-4 of Table 5 show the individual and combined contributions of our two contrastive losses. We see around a 0.5% decrease in performance when only ablating one contrastive loss, which may be due to the other contrastive loss still assisting the model in improving its learned features through relative comparisons. However, performance drops by nearly 2% when removing both contrastive losses, exhibiting that constructing triplets for contrastive learning during training is superior to simple single-label cross-entropy learning for this task.

Effectiveness of Orthogonal Loss The last row in Table 5 shows that we obtain a roughly 1% increase when using the orthogonal loss. The largest drops in performance occur from removing either the contrastive losses or view information during training, indicating that these losses contribute more directly to our learning objective than orthogonal loss.

Analysis

Visualizing Disentanglement To further exhibit disentanglement, Figure 3 shows the clustering of the learned action and view features for the 10 most common actions in the PKU-MMD cross-subject test set. A single test sample is randomly selected as a reference point (shown by ∗). We then calculate the distance between its action feature and the action features of every other sample, and do the same
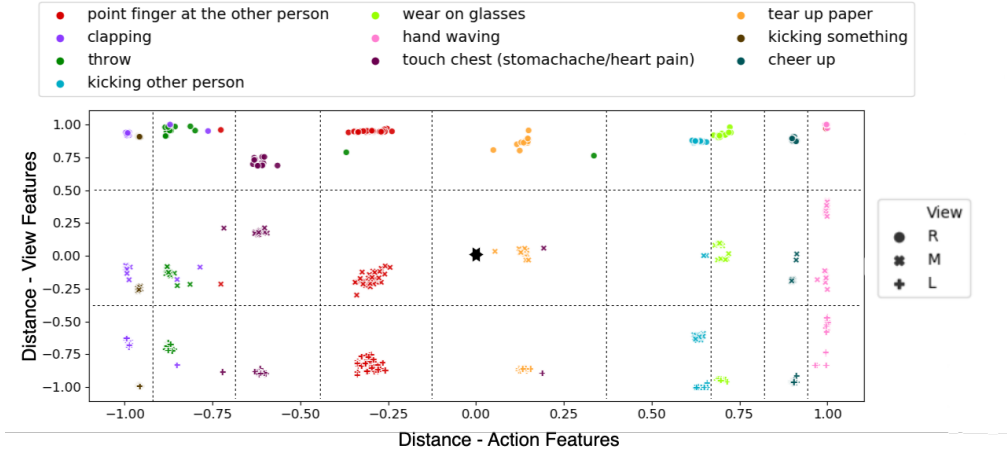
Figure 3: Visualizations of view and action features to show feature disentanglement. Each point is plotted by the distance of its respective action and view features from the reference point (∗). The learned action features form clusters by action class, and within each cluster are separated by viewpoint. Grid lines are added to help show the segregation of clusters.

$\mathcal{L}_{ace}$	$\mathcal{L}_{vce}$	$\mathcal{L}_{ac}$	$\mathcal{L}_{vc}$	$\mathcal{L}_{ortho}$	**Cross-Subject**
✓	✗	✓	✗	✓	94.0
✓	✓	✗	✓	✓	95.3
✓	✓	✓	✗	✓	95.2
✓	✓	✗	✗	✓	94.2
✓	✓	✓	✓	✗	95.0
✓	✓	✓	✓	✓	**95.8**

Table 5: PKU-MMD results when certain training losses are ablated. We achieve the highest performance with all losses.

for the view features. The x-axis denotes the distance of a given sample's action feature from our reference point's action feature, with the y-axis denoting the same for the view features. Each action class (denoted by color) clusters separately along the x-axis from other action classes in the feature space. Additionally, for each action class there are three distinct clusters along the y-axis: one for each viewpoint. The deduction from these observations is two-fold: the instances of a certain action class cluster together, indicating that our model produces action features that are distinguishable from other action classes. Secondly, instances of the same action from different viewpoints are clustered separately along the y-axis. Thus, the learned view features of our model were successfully disentangled from the action features.

Conclusion

We propose a novel transformer decoder-based architecture in tandem with two supervised contrastive losses for multi-view action recognition. By disentangling the view-relevant features from action-relevant features, we enable our model to learn action features that are robust to change in viewpoints. We show that changes in viewpoint impart perturbations on learned action features, and thus disentangling these
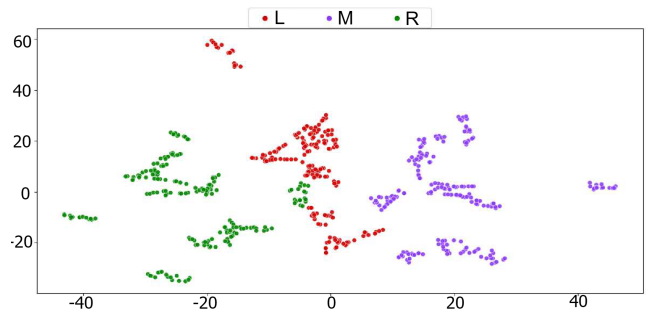

Figure 4: t-SNE of the view features from our cross-view model evaluated on the cross-subject test set. The 'left' viewpoint (shown in red) is unseen during training and is still separated from the known views during testing.

perturbations improves overall action recognition performance. Uni-modal state-of-the-art performance is attained on four large-scale, multi-view action recognition datasets.

Acknowledgements

References

Baradel, F.; Wolf, C.; Mille, J.; and Taylor, G. W. 2018. Glimpse clouds: Human activity recognition from unstruc-

tured feature points. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 469–478.

Bian, C.; Feng, W.; Meng, F.; and Wang, S. 2023. Global-local contrastive multiview representation learning for skeleton-based action recognition. *Computer Vision and Image Understanding*, 103655.

Black, J.; Ellis, T.; and Rosin, P. 2002. Multi view image surveillance and tracking. In *Workshop on Motion and Video Computing, 2002. Proceedings.*, 169–174. IEEE.

Bruce, X.; Liu, Y.; Zhang, X.; Zhong, S.-h.; and Chan, K. C. 2022. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chakraborty, B. K.; Sarma, D.; Bhuyan, M. K.; and Mac-Dorman, K. F. 2018. Review of constraints on vision-based gesture recognition for human–computer interaction. *IET Computer Vision*, 12(1): 3–15.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13359–13368.

Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; and Lu, H. 2020. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 183–192.

Cheng, Q.; Liu, Z.; Ren, Z.; Cheng, J.; and Liu, J. 2022. Spatial-Temporal Information Aggregation and Cross-Modality Interactive Learning for RGB-D-Based Human Action Recognition. *IEEE Access*, 10: 104190–104201.

Chi, H.-g.; Ha, M. H.; Chi, S.; Lee, S. W.; Huang, Q.; and Ramani, K. 2022. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20186–20196.

Das, S.; and Ryoo, M. S. 2023. ViewCLR: Learning Self-supervised Video Representation for Unseen Viewpoints. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5573–5583.

Das, S.; Sharma, S.; Dai, R.; Bremond, F.; and Thonnat, M. 2020. Vpn: Learning video-pose embedding for activities of daily living. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 72–90. Springer.

Duan, H.; Wang, J.; Chen, K.; and Lin, D. 2022. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, 7351–7354.

Elias, P.; Jan, S.; and Zezula, P. 2019. Understanding the gap between 2D and 3D skeleton-based action recognition. In *2019 IEEE International Symposium on Multimedia (ISM)*, 192–1923. IEEE.

Fei, Y.; Liu, Y.; Wei, X.; and Chen, M. 2022. O-ViT: Orthogonal Vision Transformer. *arXiv preprint arXiv:2201.12133*.

Gao, Z.; Xuan, H.-Z.; Zhang, H.; Wan, S.; and Choo, K.-K. R. 2019. Adaptive fusion and category-level dictionary learning model for multiview human action recognition. *IEEE Internet of Things Journal*, 6(6): 9280–9293.

Guo, T.; Liu, H.; Chen, Z.; Liu, M.; Wang, T.; and Ding, R. 2022. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 762–770.

Higgins, I.; Amos, D.; Pfau, D.; Racaniere, S.; Matthey, L.; Rezende, D.; and Lerchner, A. 2018. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*.

Hussain, T.; Muhammad, K.; Ding, W.; Lloret, J.; Baik, S. W.; and de Albuquerque, V. H. C. 2021. A comprehensive survey of multi-view video summarization. *Pattern Recognition*, 109: 107567.

Ismail, K. A. 2010. *Application of computer vision techniques for automated road safety analysis and traffic data collection*. Ph.D. thesis, University of British Columbia.

Ji, Y.; Yang, Y.; Shen, H. T.; and Harada, T. 2021. View-invariant action recognition via Unsupervised AttentioN Transfer (UANT). *Pattern Recognition*, 113: 107807.

Jin, X.; He, T.; Zheng, K.; Yin, Z.; Shen, X.; Huang, Z.; Feng, R.; Huang, J.; Chen, Z.; and Hua, X.-S. 2022. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14278–14287.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kong, Y.; Ding, Z.; Li, J.; and Fu, Y. 2017. Deeply learned view-invariant features for cross-view action recognition. *IEEE Transactions on Image Processing*, 26(6): 3028–3037.

Li, C.; Zhong, Q.; Xie, D.; and Pu, S. 2018a. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*.

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2018b. Unsupervised learning of view-invariant action representations. *Advances in neural information processing systems*, 31.

Lin, J.; Gan, C.; Wang, K.; and Han, S. 2020. TSM: Temporal shift module for efficient and scalable video understanding on edge devices. *IEEE transactions on pattern analysis and machine intelligence*, 44(5): 2760–2774.

Liu, A.-A.; Xu, N.; Nie, W.-Z.; Su, Y.-T.; Wong, Y.; and Kankanhalli, M. 2016. Benchmarking a multimodal and multiview and interactive dataset for human action recognition. *IEEE Transactions on cybernetics*, 47(7): 1781–1794.

Liu, C.; Hu, Y.; Li, Y.; Song, S.; and Liu, J. 2017. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10): 2684–2701.

Liu, Y.-C.; Yeh, Y.-Y.; Fu, T.-C.; Wang, S.-D.; Chiu, W.-C.; and Wang, Y.-C. F. 2018. Detach and adapt: Learning cross-domain disentangled deep representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8867–8876.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 143–152.

Piergiovanni, A.; and Ryoo, M. S. 2021. Recognizing actions in videos from unseen viewpoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4124–4132.

Shah, K.; Shah, A.; Lau, C. P.; de Melo, C. M.; and Chellappa, R. 2023. Multi-View Action Recognition Using Contrastive Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3381–3391.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2021. Adasgn: Adapting joint number and model size for efficient skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13413–13422.

Song, S.; Lan, C.; Xing, J.; Zeng, W.; and Liu, J. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Song, Y.-F.; Zhang, Z.; Shan, C.; and Wang, L. 2022. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 1474–1488.

Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; and Liu, J. 2022. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*.

Tran, L.; Yin, X.; and Liu, X. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1415–1424.

Trivedi, N.; and Ravi, S. 2023. PSUMNet: Unified Modality Part Streams are All You Need for Efficient Pose-based Action Recognition. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, 211–227. Springer.

Vyas, S.; Rawat, Y. S.; and Shah, M. 2020. Multi-view action recognition using cross-view video prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, 427–444. Springer.

Wang, D.; Ouyang, W.; Li, W.; and Xu, D. 2018a. Dividing and aggregating network for multi-view action recognition.

In *Proceedings of the European Conference on Computer Vision (ECCV)*, 451–467.

Wang, L.; Ding, Z.; Tao, Z.; Liu, Y.; and Fu, Y. 2019. Generative multi-view human action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6212–6221.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2018b. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11): 2740–2755.

Xia, L.; Chen, C.; and Aggarwal, J. 2012. View invariant human action recognition using histograms of 3D joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, 20–27. IEEE.

Zhang, A.; Chan, A.; Tay, Y.; Fu, J.; Wang, S.; Zhang, S.; Shao, H.; Yao, S.; and Lee, R. K.-W. 2021. On orthogonality constraints for transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 375–382.

Zhang, J.; Lin, L.; and Liu, J. 2023. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3427–3435.

Zhang, Z.; Tran, L.; Yin, X.; Atoum, Y.; Liu, X.; Wan, J.; and Wang, N. 2019. Gait recognition via disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4710–4719.

Zhao, L.; Wang, Y.; Zhao, J.; Yuan, L.; Sun, J. J.; Schroff, F.; Adam, H.; Peng, X.; Metaxas, D.; and Liu, T. 2021. Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12793–12802.

Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; and Kautz, J. 2019. Joint discriminative and generative learning for person re-identification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2138–2147.

Zhou, B.; Wang, P.; Wan, J.; Liang, Y.; Wang, F.; Zhang, D.; Lei, Z.; Li, H.; and Jin, R. 2022a. Decoupling and recoupling spatiotemporal representation for RGB-D-based motion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20154–20163.

Zhou, D.; Liu, Z.; Wang, J.; Wang, L.; Hu, T.; Ding, E.; and Wang, J. 2022b. Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19568–19577.