Pose-based Contrastive Learning for Domain Agnostic Activity Representations

David Schneider M. Saquib Sarfraz Alina Roitberg Rainer Stiefelhagen
Karlsruhe Institute of Technology
{firstname.lastname}@kit.edu

Abstract

While recognition accuracies of video classification models trained on conventional benchmarks are gradually saturating, recent studies raise alarm about the learned representations not generalizing well across different domains. Learning abstract concepts behind an activity instead of overfitting to the appearances and biases of a specific benchmark domain is vital for building generalizable behaviour understanding models. In this paper, we introduce Pose-based High Level View Contrasting (P-HLVC), a novel method that leverages human pose dynamics as a supervision signal aimed at learning domain-invariant activity representations. Our model learns to link image sequences to more abstract body pose information through iterative contrastive clustering and the Sinkhorn-Knopp algorithm, providing us with video representations more resistant to domain shifts. We demonstrate the effectiveness of our approach in a cross-domain action recognition setting and achieve significant improvements on the synthetic-to-real Sims4Action benchmark.[1]

1. Introduction

End-to-end deep learning facilitated a remarkable progress in the field of human activity recognition (HAR) [12, 68, 76] with impressive accuracies reported on datasets such as HMDB51 [40] or Kinetics-400 [12]. However the vast majority of published methods rely heavily on the assumption that the data used in training and testing is *independently and identically distributed (i.i.d)*. This assumption is rather naive in real-world applications, where we continuously experience *domain shifts*, *e.g.*, through changes of sensor type or placement. Recent activity recognition research provides worrying evidence that modern activity recognition frameworks are highly sensitive to changes in data distribution [57, 58] (for example an accuracy drop of $> 60\%$ is reported in [58] if the domain switches from synthetic to real data).

[1]Code: https://github.com/simplexsigil/p-hlvc

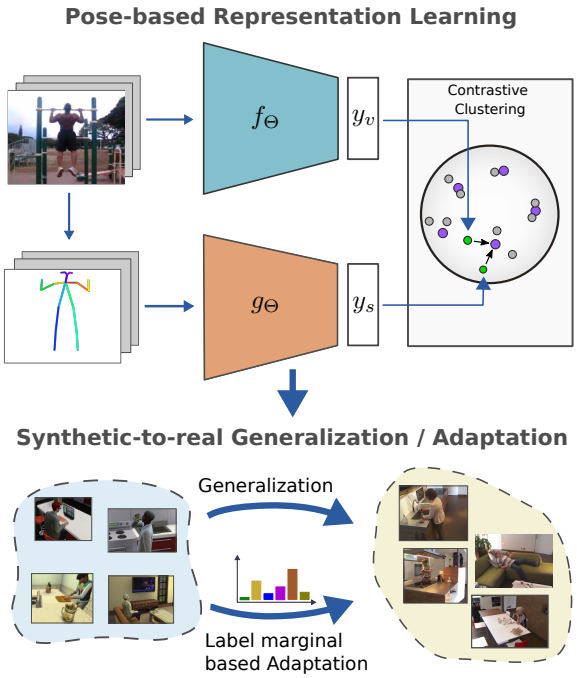

Figure 1. We pre-train a network on body pose dynamics and use the learned representations in a cross-domain manner to learn actions from synthetic data. We show that this generalizes well to real world scenarios and present an unsupervised adaptation strategy to improve results further.

Existing CNN-based activity recognition approaches [12, 68] often pre-train their models on RGB videos of large *labelled* activity recognition datasets in a *supervised* manner, after which the models are fine-tuned for the target downstream task. Categorizing activities after their appearance has changed is very difficult for such models [58], but humans handle this task without any effort. *What helps us identify an action?* While the appearance in RGB videos is strongly affected by, *e.g.*, a transition from synthetic to real data, more abstract modalities, such as body poses, have higher tolerance in this regard. Multiple frameworks successfully leveraged alternative supervision

signals computed from the videos itself (*e.g*. optical flow) in the context of self-supervised learning [5, 27], but this paradigm remains underresearched for body pose dynamics or performance under distributional shifts which is the main motivation of our work.

We seek to investigate body poses as an effective supervision signal for pre-training video representation models less susceptible to changes of data distribution and introduce Pose-based High Level View Contrasting (P-HLVC) – a new model for representation learning of human activities. A key ingredient of our approach is iterative contrastive clustering applied jointly on the videos and the body poses extracted from them. Our model learns to connect videos and the more abstract skeleton representations through the Sinkhorn-Knopp algorithm as a pre-training step, after which the video embedding model can be fine-tuned for the downstream task, leading to activity recognition models much more tolerant to domain shifts. However, the benefit of the P-HLVC approach goes beyond well-generalizable activity representations. Since the supervision signal is extracted from the data itself, our model *does not require any activity labels* during the body pose-based pre-training.

We evaluate our idea in the context of domain generalization on a synthetic-to-real activity recognition benchmark Sims4Action as well as for video retrieval on HMDB51 and UCF101, yielding a significant improvement in recognition quality in all settings. Our findings provide encouraging evidence, that modern activity recognition frameworks can benefit more from learning to connect video data and body pose sequences as part of pre-training, especially for learning domain-agnostic video representations.

To summarize, our main contributions are:

- We explore human pose dynamics as a supervision signal for learning *domain-invariant* activity representations and introduce the novel Pose-based High Level View Contrasting (P-HLVC) model.

- We conduct in-depth experiments in cross-domain human activity recognition and demonstrate clear benefits of the proposed model. Our approach considerably improves upon the state-of-the art on the *synthetic→real* activity recognition benchmark Sims4Action. Additionally, our approach does not require any category labels in the pre-training step and performs on-par with the fully supervised approaches.

- As retrieval problems also require well-generalizable feature encoders, we further validate the quality of our model in the action retrieval task on HMDB51 and UCF101, yielding state-of-the-art results.

2. Related work

2.1. Representation learning

Representation learning without manual annotations is well explored, due to its applicability for training on very large unlabelled datasets. Recent self-supervised methods typically either train to detect low or high level instance transformations [5, 35] or they split the information contained in a sample on the instance level which may include separating color channels or clips along the time dimension and train by matching representations of the separated data or by reconstructing the information [22, 25, 26, 33, 51, 75, 76]. [27] make use of paired video and optical flow sequences to mine positive class samples in an alternating way. Contrastive learning is often applied on multiple generated views of existing data, for example after applying image transformations [14, 28, 62] or as multi-modal contrastive learning [1, 38, 43, 52, 53]. Contrastive clustering based losses have been used for image representation learning [3, 11] as well as recently for learning representations from video [2]. The research most similar to ours is presumably the concurrent work of Rai *et al*. [56]. However, while Rai *et al*. [56] also briefly consider body poses, their experimental focus is clearly put on image-based modalities.

2.2. Action recognition with body poses

With the availability of good body pose estimation methods from video [10], performing action recognition from pose has been explored in many different works for example with Recurrent Neural Networks [44, 74], Graph Convolutional Networks [60, 71], or with CNN based methods on generated pseudo images [7, 8, 16]. Some works use body poses and RGB sequences effectively in combination by using pose information as an attention mechanism, for example [19] or [18]. However, these works do not aim at learning representations contrastively but rather leverage pose or pose feature extractors as an additional source of information for categorical supervised classification.

2.3. Synthetic datasets and domain adaptation

Until recently, very few datasets targeted action recognition from synthetic data, significant progress has been made over the past years. [20, 46, 65] and [29] evaluate the usage of synthetic data to augment real data during training, [54] use it to learn action compositions and [58] intend to learn human actions from virtual data only. Most unsupervised domain adaptation frameworks aim to match within-network feature distributions of the source and target domain, for example by leveraging a domain adversarial loss [24, 31]. Newer publications extend this by combining it with attention to align multiple temporal cross-domain features [13], with a self-supervised learning loss which is

applied on both domains, source and target [15] or the extension to multiple modalities [49].

3. High level view contrasting

In this work we explore contrastive multi-view representation learning limited to the video modality paired with body poses generated from video sequences. Human body poses can be extracted using off-the-shelf detectors which are trained in a self-supervised way, for example [30,36,64], but results tend to be better with supervised models like [10]. Note, that we do not need the pose estimator to be trained on our pre-training dataset, which allows us to pre-train on unlimited datasets without human labelling effort. We represent body joint movements with the SkeleMotion representation of [8], which calculates the magnitude and direction of joint movements and arranges the results in an image like structure which is well suited to be interpreted with a very lightweight CNN. One advantage of body joints instead of arbitrary points of interest or optical flow maps is the inherent semantically meaningful structure of the data. This allows for a smaller representation in comparison to unstructured data like optical flow maps.

Our pre-training method draws inspiration from SvAW [11]. They use a single CNN encoder to calculate representations for multiple augmented 2D images and then project these representations onto movable cluster centers which are distributed on the unit sphere. Since this would lead to trivial solutions, they use the work of [3] and apply the Sinkhorn-Knopp algorithm to match a calculated target cluster assignment of paired images instead of the projected cluster assignment itself.

Instead of applying low level data augmentations with the same representation encoder for all augmented views, we use generated body poses as a high level view and apply two different representation encoders f_Θ and g_Θ to work with the different data sources. A detailed overview of our architecture is provided in the supplementary.

3.1. Representation space

Video Embeddings For a video $v \in \mathbb{R}^{L\times W\times H\times 3}$ with L being the length of the video (Kinetics-400: 10 seconds, 30 fps sampled, $L = 300$), W and H being the height and width and the last dimension representing the RGB space, we sample sub-sequences $x_v \in \mathbb{R}^{T\times H\times W\times 3}$ with T being the desired clip length, randomly on each epoch. After applying randomly chosen data transformations like cropping, rotations or color changes (consistent over the clip length), we use a 3D CNN video encoder f with parameters Θ to attain a normalized representation vector $f_\Theta(x_v) = y_v \in \mathbb{R}^m, \|y_v\| = 1$ with m being the representation vector dimensionality. We refer to the space $\mathbb{R}^m$ containing the representation vectors as the *Representation Space*. The encoder f is either a slightly adapted version of the R-2D3D network [25], the S3D network [68] or MoViNet A2 [37], however, f can easily be exchanged with other architectures.

Body pose-time embeddings Given a body pose sequence $x_b \in \mathbb{R}^{J\times L\times 3}$ with J being the number of body joints, L being the time steps in the video and the last dimension describing the positions in three dimensional space, we compute a SkeleMotion representation using the work of [8] and then select clip length crops in order to achieve a representation $x_s \in \mathbb{R}^{T\times I\times 6}$ which encodes the orientation and magnitude of body joint movements from consecutive frames. We use these representations as input to compute body pose-time embeddings $g_\Theta(x_s) = y_s \in \mathbb{R}^m, \|y_s\| = 1$ with our body pose dynamic representation encoder g_Θ, likewise to the generation of the video clip representations. The encoder g is a custom lightweight and simple CNN, similar to the one proposed in [8]. A SkeleMotion image of size $32 \times 49 \times 6$ which is paired with a video clip of size $32 \times 128 \times 128 \times 3$ increases the size of the input data per sample by less than 1% in comparison to double the size when using an additional augmented view of the video clip. This effect is even more significant with larger input clips of size $224 \times 224 \times 3$, since the size of the SkeleMotion representations does not increase with image size. Doing contrastive learning on multiple views which consist of paired image data or optical flow is computationally much more expensive.

3.2. Assignment space

We define a set C of n cluster center vectors $C = [c_1, .., c_n]^T \in \mathbb{R}^{n\times m}, \|c_i\| = 1, i \in \{1..n\}$ in the representation space which are influenced by optimizing the loss function. For each vector y in representation space we generate cluster assignment vectors $z \in \mathbb{R}^n_{[-1,1]}$ where $z = Cy$. We refer to $\mathbb{R}^n_{[-1,1]}$ as the assignment space. Finally, we generate assignment prediction vectors $p = \mathrm{softmax}(z)$ which contain probability values for each cluster.

It is not clear how many cluster centers are effective. A small number of learnable clusters may impact performance, too many clusters at the start can prevent successful training. This can be handled by growing the number of clusters during training, as explained in Section 3.4.

To enforce an equal partitioning of the data to the clusters, we use the Sinkhorn-Knopp algorithm to calculate the target assignments $Q = [q_1, .., q_n]^T = \mathrm{sinkhorn}(exp(\frac{[z_1,..,z_n]^T}{\epsilon}) \in \mathbb{R}^{nxm}$ with ϵ being a parameter to influence the smoothness of the target assignment. Using a small ϵ is crucial to distribute the data to multiple clusters successfully. Empirically, we found $\epsilon < 0.012$ to be sufficient for preventing the backbone networks from converging to a trivial solution.

3.3. The agreement accentuating CE loss

Although our approach was tested to work with the original SwAV-loss from [11], it requires a careful choice of hyper parameters to prevent it from converging to a trivial solution, especially at the beginning of training. It is not surprising that training in this setting is more difficult than for [11] in the image domain, since we apply a two-stream architecture and our two types of input data have very different structures and semantics. For this reason we propose a new loss function that is better suited for the task of contrastive clustering in the video domain.

For a cluster assignment prediction p and its paired target cluster assignment q (from another view), [11] define the loss l_{SwAV} in Eq. 1:

$$l_{\text{SwAV}}(p,q) = -\sum_k q^{(k)} \log p^{(k)} \tag{1}$$

Instead, we propose the loss l_{AXent}:

$$l_{\text{AXent}}(p,q) = -\sum_k q^{(k)} \log p^{(k)} + (1-q^{(k)}) \log(1-p^{(k)}) \tag{2}$$

We call this loss the *Agreement Accentuating Cross-Entropy Loss* (l_{AXent}). On a first glance it is similar to the often used definition of the binary cross-entropy loss. However, these are not the same functions. The binary cross-entropy loss expects a dual class problem with $q \in \{0,1\}$ which means that in any case only one of the two terms will be non-zero. AXent on the other hand is defined for $q \in [0,1]$. This is a linear combination of the two edge cases which make up binary cross-entropy and perfectly well suited for a problem, where q is a continuous variable indicating the attraction to a desired class and the loss is minimized by varying p. In a system where $q \sim p$, this loss function accentuates the agreement of these two variables either towards $(p,q) = (0,0)$ or towards $(p,q) = (1,1)$.

In a batch of N samples, each sample might be augmented multiple times to generate multiple video views V with $|V| = K_v$ and multiple body pose tracks S_i with $|S_i| = K_s$. While K_v is fixed during training, K_s may be different for every sample and is only limited by a maximal number of the allowed body tracks. We make use of all $V \times S_i, i \in \{1..N\}$ combinations per sample. l_{AXent} is used twice, once on the body pose assignment prediction p_i^s and video target assignment q_i^v and once on the video assignment prediction p_i^v and body pose target assignment q_i^s.

$$l = -\frac{1}{N}\sum_{i=1}^{N}\sum_{s\in S_i}\sum_{v\in V}\frac{l_{\text{AXent}}(p_i^s, q_i^v) + l_{\text{AXent}}(p_i^v, q_i^s)}{2} \tag{3}$$

We ablate the AXent-loss function in table 1 and provide a further analysis of the differences between the AXent-loss and the SwAV-loss in the supplementary.

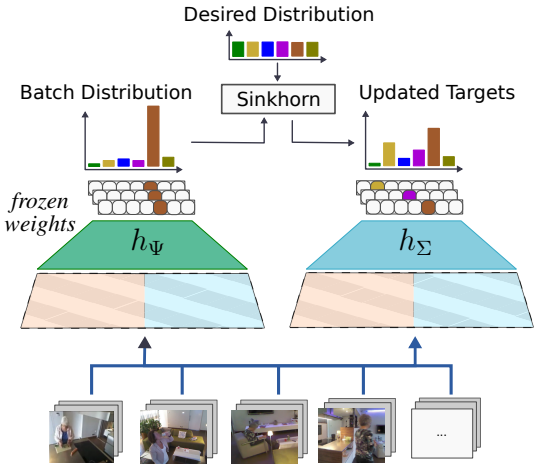

Figure 2. Overview of our domain adaptation method. We evaluate the batch distribution of label predictions and use the Sinkhorn-Knopp algorithm to calculate target predictions which conform to a desired distribution.

3.4. Iteratively growing the assignment space

Empirically, we found that there is a certain relation between the dimensionality of the assignment space and the number of cluster centers, as using significantly more or less cluster centers increases the probability for the algorithm to converge to a trivial solution which does not transfer well. We enabled the ability to train on very large numbers of clusters with an iterative cluster splitting algorithm. In order to distribute the data more evenly, we sum the cluster assignments for each epoch and select the clusters with most assignments as splitting candidates d_i. New cluster centers are then introduced as $\hat{d}_i = d_i + d_\epsilon$ with $d_\epsilon \in \mathbb{R}^n_{[-\epsilon,+\epsilon]}$. We then rely on the sinkhorn algorithm and gradient updates to push the cluster duplicates apart and to distribute them more evenly on the unit sphere.

3.5. Unsupervised domain adaptation

We leverage our pre-trained architecture by evaluating it on unsupervised domain adaptation from synthetic to real world data and append a classifier h_Ψ on top, which is then trained on the source domain. To further improve our transfer performance, we present an unsupervised self-calibration technique which performs teacher-student training on the target domain, an overview is provided in Figure 2. We assume a plausible label distribution and enforce it during batch-wise training on the target dataset by phrasing the assignment of labels to classes as an optimal transport problem from a uniform marginal to the assumed prediction marginal with the class scores per sample forming a

Modalities	Loss	UCF		HMDB	
		Top 1	Top 5	Top 1	Top 5
Video + Video	AXent	58.1	85.5	30.0	66.0
Video + Pose	InfoNCE	68.0	92.5	39.8	74.8
Video + Pose	SwAV	81.0	95.0	52.8	81.2
Video + Pose	AXent	83.0	96.3	55.1	82.2

Table 1. Transfer learning comparison of combinations of different loss functions and modalities.

cost matrix. Similar to our pre-training procedure, we can use the Sinkhorn-Algorithm to solve this problem and receive new predictions which follow the assumed distribution more closely. A student network h_Σ which was initialized with the teacher weights Ψ is then trained on the target dataset to predict the teacher's calibrated predictions. While distribution matching is one of the foundational principles of domain adaptation [4, 45, 77], most approaches perform this on the feature level. This does not apply to our architecture, since our backbone is pre-trained on real world examples and frozen during fine-tuning on the synthetic domain. Despite a similar underlying intuition, the usage of the Sinkhorn-Knopp algorithm, our teacher-student architecture as well as the focus on the label distribution rather than the within network feature distribution separates our technique from previous approaches.

A reasonable assumption for the desired label distribution is the label distribution of the source dataset, additionally our method can be transformed into weakly supervised domain adaptation by providing an oracle based distribution or sampling a small part of the target domain samples to generate a distribution estimate.

4. Experiments

We use Kinetics-400 [34] (videos sampled at 30 frames per second) as our pre-training dataset and HMDB51 [40], UCF101 [61] and Sims4Action [58] as the downstream datasets. We do not use any labels of Kinetics-400. Evaluation after performing transfer learning on Sims4Action is perfomed on Toyota Smarthome [17] as described in [58]. We also use the Kinetics-Skeleton dataset [71] containing pose sequences for Kinetics-400. We note that poses can also be extracted using self-supervised detectors, e.g. [30, 36, 64]. The Kinetics-Skeleton dataset made use of an off-the-shelf pose estimator which was trained on data unrelated to Kinetics-400, preventing any manually labelled supervisory signal from Kinetics itself to influence our results. This is similar to [43] who make use of supervised speech recognition models to generate text from audio for multi-modal representation learning.

Our dataset contains 122k video samples after filtering videos which do not have a correspondence in the Kinetics-Skeleton dataset, while the original training set contains 224K videos. It is interesting to observe that even with *half of the training examples*, our approach performs on par with approaches trained on the full K400 data. This further signifies the potential and effectiveness of our approach. Unless noted otherwise, evalutation is performed on the video feature encoder f.

4.1. Ablation studies

First, we validate three key ingredients of our approach: (1) contrastive clustering as learning mechanism, (2) the usage of body pose as a complementing view and (3) the introduction of the BXent loss. We compare our contrastive clustering method with "vanilla" contrastive learning using the InfoNCE loss [14] while keeping the same backbone architectures and hyper parameters for both methods. To evaluate the feasibility of body poses as a learning signal, we train a network on two differently augmented video views sampled from two random locations within a video instead of combining video and pose for 200 epochs and using frames of size 128×128. The results are listed in Table 1 where we compare downstream transfer performance to HMDB51 and UCF101 as well as in Table 5 where the *InfoNCE* experiment is referred to as P-HLVC NCE and the video-video contrastive learning experiment is referred to as V-HLVC. Pose-based contrastive learning results in strong representations for action video retrieval (Table 5), while video-video contrastive learning does not reach this performance, despite extensive augmentations with temporal shifts between the matching views. Note, that training with two video views requires roughly the double amount of resources (GPU, RAM and time). Pre-training with the AXent loss results in measurably improved transfer learning performance compared to the original SwAV loss, improving top-1 performance on HMDB51 by 3.1% and on UCF101 by 2%.

4.2. State-of-the-art comparison

We compare our model to state-of-the-art approaches in the context of cross-domain synthetic-to-real recognition as well as on k-Nearest Neighbour (K-NN) action retrieval. The variations in architectures, image sizes and clip lengths for existing approaches make it difficult to provide a one to one comparison with all existing methods. For this reason, we list P-HLVC results for three different lightweight architectures, the R2D3D architecture as used in [25, 26], the S3D architecture as used in [5, 27, 48, 58, 68] and the recently presented MoViNet architecture [37] in its A2 setting. Detailed information about our training procedure and architecture as well as transfer learning results on HMDB51 and UCF101 are provided in the supplementary.

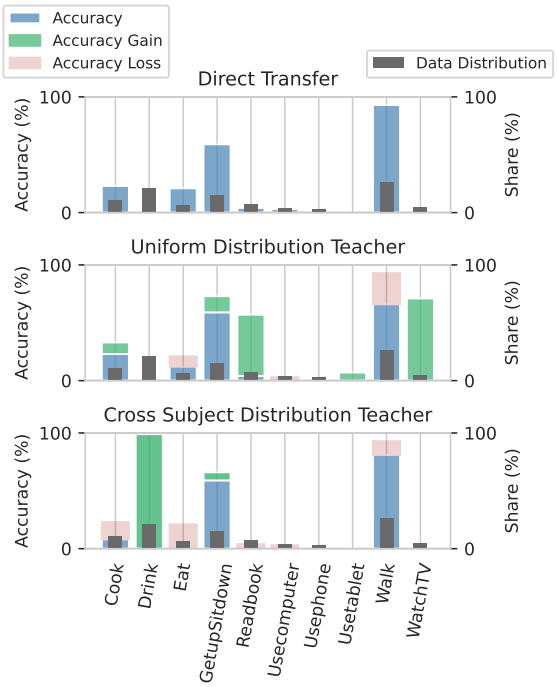

Figure 3. Comparison of direct transfer (top) with unsupervised distribution training assuming the source label distribution (uniform) or using the target cross-subject train label distribution.

4.2.1 Cross-domain action classification

We evaluate the impact of the proposed pre-training technique on model generalization to new domains using the Sims4Action testbed [58] for *synthetic→real* transfer in Table 2. Additionally, we test our domain adaptation method described in Section 3.5 and present the results in Table 3.

After our pose-supervised representation learning, we freeze all pre-trained weights during the transfer learning on synthetic Sims4Action videos, except for one experiment marked with †† to illustrate the effects of end-to-end fine-tuning. The evaluation takes place on *real* data obtained from Toyota Smarthome [17]. Sims4Action is further referred to as the *source dataset* and Toyota Smarthome as the *target dataset*. We follow evaluation procedure of [58] and predict a label for the middle 90 frame chunk per video. Additionally, we evaluate on the full video, using averaged ten-crop predictions in order to provide comparable results for approaches which predict an action based on longer video clips. We reprt the accuracy as well as the mean Per-Class Accuracy (mPCA / balanced accuracy).

Pre-training significantly improves generalization performance (see Table 2). We find that the best results are achieved by keeping the trainable classification head as shallow as possible, the models with three fully-connected layers or full end-to-end fine-tuning perform worse, which

Method	Pre-training	Full Video		Mid-Chunk	
		Acc	mPCA	Acc	mPCA
Domain Generalization (No Pre-Training)					
S3D [58]		18.5	13.4	20.0	12.4
Ours$_V$		13.4	10.0	13.0	9.1
Ours$_B$		12.2	14.0	12.0	12.7
Ours$_{V+B}$		11.8	13.3	12.4	13.2
Domain Generalization (Pre-trained)					
S3D [58]	K400	36.0	27.3	34.1	23.2
TA^{3}N [13]	IN	12.42	13.61		-
APN [73]	IN	18.1	19.7		-
VideoDG [73]	IN	19.6	23.6		-
Ours$_V$	|	38.8	21.3	35.6	19.8
Ours$_V$†	|	37.85	20.6	35.2	19.4
Ours$_V$††	KS	20.9	18.9	20.7	19.3
Ours$_B$	|	24.8	19.6	24.7	18.8
Ours$_{V+B}$	|	**40.4**	**29.0**	**38.3**	**28.0**

V Video **B** SkeleMotion **V+B** Concatenated Features
K400 Kinetics-400 **IN** ImageNet **KS** Kinetics-Skeleton
† Classifier using three linear layers instead of one.
†† End-to-end fine-tuning.

Table 2. Domain generalization results from Sims4Action to Toyota Smarthome with evaluation on the cross-subject test set following [58]. Our approach only needs a dataset with paired body skeleton sequences for pre-training.

Method	Target Superv.	Full Video		Mid-Chunk	
		Acc	mPCA	Acc	mPCA
Unsupervised Domain Adaptation					
TA^{3}N [13]		8.8	12.7		-
Ours		40.6	**31.3**	36.0	**28.1**
Ours†		39.4	29.3	36.3	27.1
(Weakly) Supervised Domain Adaptation					
TA^{3}N [13]	Labels	33.1	13.4		-
Ours	Dist.	**53.4**	25.5	**49.9**	24.3
Ours†	Dist.	53.2	26.0	49.3	25.5

† Classifier using three linear layers instead of one.

Table 3. Unsupervised and supervised domain transfer from Sims4Action to Toyota Smarthome with evaluation on the cross-subject test set. Our domain adaptation technique is weakly supervised, since we only make use of the label distribution in contrast to using the labels themselves.

Method	ROSE	UCF	HMDB
P-HLVC (S3D)	✗	83.2	54.4
P-HLVC (S3D)	✓	59.5 (-23.7)	44.5 (-9.9)

Table 4. Transfer learning results on the original HMDB51 and UCF101 test splits and on the ROSE challenge benchmark.

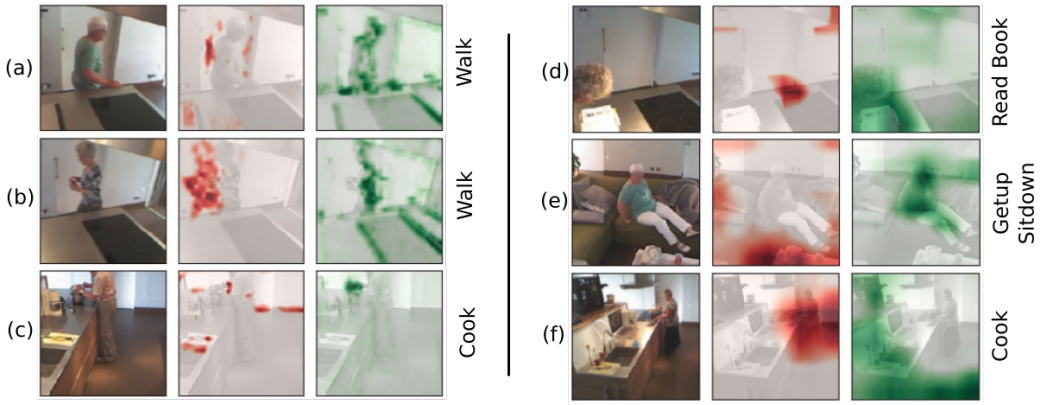

Figure 4. Grad-CAM [59] results on Toyota Smarthome. The columns depict the original image as well as negative attributions and positive attributions shown as heatmaps. (a)-(c) and (d)-(f) show Grad-CAM being applied at convolutional layers three and nine, respectively. It is clearly visible, how our method focuses on the human body as an indicator to determine an action.

we attribute to the deterioration of well-transferable backbone weights and overfitting. This supports our assumption that a pre-trained feature encoder for domain generalization should be as descriptive of human actions as possible, since a more general pre-training might require a more sophisticated multi-layer classification head.

Apart from the comparison to [58], we also provide results for recent domain generalization and unsupervised domain adaptation frameworks with publicly available implementations, VideoDG [73] (including APN as baseline) and TA^{3}N [13]. [58] provide results for the S3D architecture being pre-trained in a fully-supervised manner on Kinetics-400. This is a challenging comparison, since the supervised pre-training on Kinetics-400 can learn many of the relevant action classes on real data and only has to recognize their label assignment in the synthetic domain, while our own approach faces the concept of actions in the synthetic domain for the first time. We consider this an important scenario since the usage of synthetic data is mainly justified by avoiding the need for annotated real-world action datasets. Despite solving a harder task in this sense, our combined video and body pose backbones which were pre-trained without action labels outperform action based Kinetics-400 pre-training on all metrics.

In Table 3 our domain adaptation technique is compared with the results of [13]. We consider different calibration strategies by either using the source dataset label distribution (unsupervised) or using additional information about the label distribution in the cross-subject train set of Toyota Smarthome (weakly supervised). On unsupervised domain adaptation we achieve an accuracy of 40.6% outperforming randomly initialized domain generalization (Table 2) by 27%. The framework of [13] allows for using supervised domain adaptation, we compare these results with our work by making use of the target set label marginal for our calibration method instead of the source marginal and outperform them by 20.3% on accuracy. Unsurprisingly, assuming the uniform label distribution of Sims4Action works better for mPCA, as the metric treats every category as equal and the target dataset label marginal maximises standard accuracy. We list Figure 3 to visualize the effects of our calibration strategies on our video model.

4.2.2 Action video retrieval

Next, we evaluate the representations generated with our body pose-driven pre-training approach on HMDB51 and UCF101 action retrieval. We sample 10 clips of all videos in the test and train set of these datasets and average these clip representations per sample. We did not apply any augmentations apart from random cropping.
As commonly performed, the test set sample representations are used to query the classes of the train set samples by evaluating their representation similarity and the Recall@K is reported based on the top k returned results. The results of this experiment are listed in Table 5 where we compare with other representation learning methods which pre-train on another dataset than the retrieval set, a setting which requires good generalizability of the feature encoder. Our models show very good performance, improving over CoCLR [27] on all datasets or ViCC on HMDB51 which share the same S3D backbone network.

4.2.3 ROSE Challenge

In order to evaluate the robustness of our pose-supervised pre-training strategy, we list our S3D fine-tuning results on HMDB51 and UCF101 in Table 4 and compare them to the 2022 Robustness in Sequential Data challenge (ROSE). The performance losses are substantial, but we find that despite the heavy alterations on the input data, P-HLVC still shows a certain robustness, even outperforming older approaches

Method \ k	1	5	10	20
HMDB51 (UCF101 Pre-Training)				
VCOP [70]	7.6	22.9	34.4	48.8
VCP [47]	7.6	24.4	36.3	53.6
MemDPC (R2D3D) [26]	7.7	25.7	40.6	57.7
PRP [72]	10.5	27.2	40.4	56.2
Pace [67]	12.9	31.6	43.2	58.0
CoCLR (S3D) [27]	23.2	43.2	53.5	65.5
ViCC (S3D) [63]	25.5	49.6	61.9	72.5
HMDB51 (Kinetics Pre-Training)				
CtP (R3D-18) [66]	11.8	30.1	-	-
VCLR (R2D-50) [39]	35.2	58.4	68.8	79.8
P-HLVC (R2D3D)	8.0	27.1	41.3	58.6
V-HLVC (S3D)	9.6	25.1	38.0	53.6
P-HLVC NCE (S3D)	24.5	49.3	62.4	74.7
P-HLVC (S3D)	27.3	53.7	66.3	79.1
P-HLVC (MVN)	29.7	56.5	69.20	80.0
UCF101 (UCF101 Pre-Training)				
VCOP [70]	14.1	30.3	40.0	51.1
VCP [47]	19.9	33.7	42.0	50.5
Jigsaw [50]	19.7	28.5	33.5	40.0
OPN [41]	19.9	28.7	34.0	40.6
MemDPC (R2D3D) [26]	20.2	40.4	52.4	64.7
PRP [72]	23.2	38.1	46.0	55.7
Buchler [6]	25.7	36.2	42.2	49.2
Pace [67]	25.6	42.7	51.3	61.3
CoCLR (S3D) [27]	53.3	69.4	76.6	82.0
ViCC (S3D) [63]	62.1	77.1	83.7	87.9
UCF101 (Kinetics Pre-Training)				
SpeedNet (S3D-G) [5]	13.0	28.1	37.5	49.5
TempTrans [32]	26.1	48.5	59.1	69.6
CtP (R3D-18) [66]	29.0	47.3	-	-
VCLR (R2D-50) [39]	70.6	80.1	86.3	90.7
P-HLVC (R2D3D)	14.0	34.4	46.7	61.1
V-HLVC (S3D)	19.1	35.5	45.0	55.8
P-HLVC NCE (S3D)	37.4	59.2	69.9	79.4
P-HLVC (S3D)	39.7	64.0	73.5	81.9
P-HLVC (MVN)	53.7	74.2	82.0	88.2

Table 5. Video retrieval on HMDB and UCF in %. If a backbone is used by mutiple approaches, best results are underlined.

like [25] on the HMDB51 test set. A full transfer-learning comparison table is listed in the supplementary.

5. Qualitative analysis

Next, we showcase the image regions driving the decision of our model using Grad-CAM [59]. Figure 4 shows alpha blended heat maps of negative and positive attribution to the ground truth prediction, respectively. Samples (a)-(c) analyze the shallow third convolutional layer of S3D, while samples (d)-(f) are the result of applying Grad-CAM after the convolutional layer number nine. We refer to the supplementary for more details on the implementation as well as further examples. The results show how our pre-training forces the network to focus on the human body, for both, positive and negative decisions. Samples (a) to (c) demonstrate, that this is not only the case for high level layers, but rather a property which is already present in low level layers. In sample (b), it is visible, that the body provides contradicting information to the network and in sample (f) it appears that the action class "cook" might actually be a result which is inferred from the environment, rather than the body itself.

6. Conclusion

P-HLVC is a new approach for representation learning from automatically generated body poses aimed to improve domain generalization of synthetic-to-real action classification. Our approach does not require any category labels and can therefore be utilized on large unlabelled video datasets. We demonstrated the quality of our feature encoder by performing ablation experiments and presenting state-of-the-art results on action retrieval as well as the difficult Sims4Action domain generalization benchmark. We further improved these results by presenting a simple but very effective unsupervised domain adaptation technique which is complementary to existing adversarial loss approaches. Our experiments demonstrate the potential of body poses as an effective domain-agnostic supervisory signal.

Broader impact and limitations P-HLVC was developed to improve assistance applications for activities of daily living. For such scenarios, the collection of datasets can be privacy infringing and great care has to be taken to protect the right of informational self-determination. With our work we hope to drive research to improve applicability of synthetic datasets which offer relieve for such problems and additionally provide opportunities to counter dataset biases for example by simulating generational or ethnical diversity. This field of research remains limited by strong domain and performance gaps which prevent synthetic datasets from being used commonly for such applications. We believe this gap to shrink with the development of realistic simulation engines, driven by the gaming industry.

Although action recognition frameworks raise ethical questions since they might be applied for surveillance or military applications, we believe that the possible merits of our specific research direction outweigh these general considerations and we hope to propel human-assistive research with P-HLVC as well as with our future efforts.

Acknowledgements This work was supported by the JuBot project which was made possible by funding from the Carl-Zeiss-Foundation.

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 15

[2] Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *arXiv preprint arXiv:2006.13662*, 2020. 2, 15

[3] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 3

[4] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007. 5

[5] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T. Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 5, 8, 14, 15

[6] Uta Buchler, Biagio Brattoli, and Bjorn Ommer. Improving spatiotemporal self-supervision by deep reinforcement learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 770–786, 2018. 8

[7] Carlos Caetano, François Brémond, and William Robson Schwartz. Skeleton image representation for 3d action recognition based on tree structure and reference joints. In *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 16–23. IEEE, 2019. 2

[8] Carlos Caetano, Jessica Sena, François Brémond, Jefersson A Dos Santos, and William Robson Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019. 2, 3

[9] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 12

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2, 3

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 3, 4

[12] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 1

[13] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Woo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 6, 7, 12

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 5

[15] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *European Conference on Computer Vision*, pages 678–695. Springer, 2020. 3

[16] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7024–7033, 2018. 2

[17] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 5, 6, 12

[18] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *arXiv preprint arXiv:2105.08141*, 2021. 2

[19] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *European Conference on Computer Vision*, pages 72–90. Springer, 2020. 2

[20] César Roberto de Souza12, Adrien Gaidon, Yohann Cabon, and Antonio Manuel López. Procedural generation of videos to train deep action recognition networks. 2017. 2

[21] Ali Diba, Vivek Sharma, Rainer Stiefelhagen, and Luc Van Gool. Weakly supervised object discovery by generative adversarial & ranking networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 601–610. Computer Vision Foundation / IEEE, 2019. 15

[22] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[23] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 12

[24] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2

[25] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Workshop on Large Scale Holistic Video Understanding, ICCV*, 2019. 2, 3, 5, 8, 14, 15

[26] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *ECCV*, 2020. 2, 5, 8, 14, 15

[27] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised Co-training for Video Representation Learning. (NeurIPS), 2020. 2, 5, 7, 8, 14, 15

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[29] Hochul Hwang, Cheongjae Jang, Geonwoo Park, Junghyun Cho, and Ig-Jae Kim. Eldersim: A synthetic data generation platform for human action recognition in eldercare applications, 2020. 2, 12

[30] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8787–8797, 2020. 3, 5

[31] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *BMVC*, volume 2, page 5, 2018. 2

[32] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 425–442. Springer, 2020. 8

[33] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *ArXiv*, abs/1811.11387, 2018. 2, 15

[34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 5

[35] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8545–8552, 2019. 2, 15

[36] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 3, 5

[37] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16030, 2021. 3, 5

[38] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7763–7774. Curran Associates, Inc., 2018. 2, 15

[39] Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Sören Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3195–3204, 2021. 8, 15

[40] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 1, 5

[41] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 8

[42] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 12

[43] Tianhao Li and Limin Wang. Learning spatiotemporal features via video and text pair discrimination, 2020. 2, 5, 15

[44] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016. 2

[45] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 5

[46] Dennis Ludl, Thomas Gulde, and Cristóbal Curio. Enhancing data-driven algorithms for human pose estimation and action recognition through simulation. *IEEE transactions on intelligent transportation systems*, 21(9):3990–3999, 2020. 2

[47] Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. *arXiv preprint arXiv:2001.00294*, 2020. 8

[48] Kyle Min and Jason J Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2394–2403, 2019. 5

[49] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 122–132, 2020. 3

[50] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016. 8

[51] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[52] Mandela Patrick, Y. Asano, Ruth Fong, João F. Henriques, G. Zweig, and A. Vedaldi. Multi-modal self-supervision from generalized data transformations. *ArXiv*, abs/2003.04298, 2020. 2, 15

[53] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Evolving losses for unsupervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, pages 133–142, 2020. 2, 15

[54] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018. 2

[55] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge J. Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *CoRR*, abs/2008.03800, 2020. 14, 15

[56] Nishant Rai, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Cocon: Cooperative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3384–3393, 2021. 2, 15

[57] Simon Reiß, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. Deep classification-driven domain adaptation for cross-modal driver behavior recognition. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1042–1047. IEEE, 2020. 1

[58] Alina Roitberg, David Schneider, Aulia Djamal, Constantin Seibold, Simon Reiß, and Rainer Stiefelhagen. Let's play for action: Recognizing activities of daily living by learning from life simulation video games. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021. 1, 2, 5, 6, 7, 12

[59] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7, 8

[60] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019. 2

[61] Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, 12 2012. 5

[62] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning, 2020. 2

[63] Martine Toering, Ioannis Gatopoulos, Maarten Stol, and Vincent Tao Hu. Self-supervised video representation learning with cross-stream prototypical contrasting. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 8, 15

[64] Hsiao-Yu Fish Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. *arXiv preprint arXiv:1712.01337*, 2017. 3, 5

[65] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7):2264–2287, 2021. 2

[66] Guangting Wang, Yizhou Zhou, Chong Luo, Wenxuan Xie, Wenjun Zeng, and Zhiwei Xiong. Unsupervised visual representation learning by tracking patches in video. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2021. 8, 15

[67] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European conference on computer vision*, pages 504–521. Springer, 2020. 8

[68] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018. 1, 3, 5

[69] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018. 12

[70] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 8

[71] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 2, 5

[72] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6548–6557, 2020. 8

[73] Zhiyu Yao, Yunbo Wang, Xingqiang Du, Mingsheng Long, and Jianmin Wang. Adversarial pyramid network for video domain generalization. *arXiv preprint arXiv:1912.03716*, 2019. 6, 7

[74] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2117–2126, 2017. 2

[75] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[76] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017. 1, 2

[77] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019. 5