2021 IEEE/CVF International Conference on Computer Vision (ICCV) | 978-1-6654-2812-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCV48922.2021.00912

Video Pose Distillation for Few-Shot, Fine-Grained Sports Action Recognition

James Hong[1] Matthew Fisher[2] Michaël Gharbi[2] Kayvon Fatahalian[1]
[1]Stanford University [2]Adobe Research

Abstract

Human pose is a useful feature for fine-grained sports action understanding. However, pose estimators are often unreliable when run on sports video due to domain shift and factors such as motion blur and occlusions. This leads to poor accuracy when downstream tasks, such as action recognition, depend on pose. End-to-end learning circumvents pose, but requires more labels to generalize.

We introduce Video Pose Distillation (VPD), a weakly-supervised technique to learn features for new video domains, such as individual sports that challenge pose estimation. Under VPD, a student network learns to extract robust pose features from RGB frames in the sports video, such that, whenever pose is considered reliable, the features match the output of a pretrained teacher pose detector. Our strategy retains the best of both pose and end-to-end worlds, exploiting the rich visual patterns in raw video frames, while learning features that agree with the athletes' pose and motion in the target video domain to avoid overfitting to patterns unrelated to athletes' motion.

VPD features improve performance on few-shot, fine-grained action recognition, retrieval, and detection tasks in four real-world sports video datasets, without requiring additional ground-truth pose annotations.

1. Introduction

Analyzing sports video requires robust algorithms to automate fine-grained action recognition, retrieval, and detection in large-scale video collections. Human pose is a useful feature when sports are centered around people.

State-of-the-art skeleton-based deep learning techniques for action recognition [31, 57] rely on accurate 2D pose detection to extract the athletes' motion, but the best pose detectors [45, 54] routinely fail on fast-paced sports video with complex blur and occlusions, often in frames crucial to the action (Figure 1). To circumvent these issues, end-to-end learned models operate directly on the video stream [7, 13, 28, 43, 51, 62]. However, because they consume pixel instead of pose inputs, when trained with few labels, they tend to latch onto specific visual patterns [9, 52] rather than the fine-grained motion (e.g., an athlete's clothes or the presence of a ball). As a result, prior pose and end-to-end methods often generalize poorly on fine-grained tasks in challenging sports video, when labels are scarce. While collecting large datasets with fine action and pose annotations is possible, doing so for each new sport does not scale.

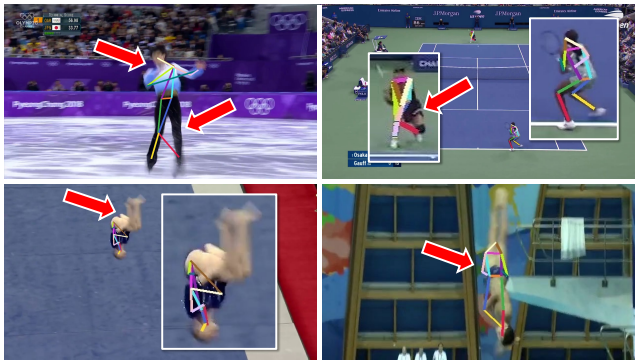

Figure 1: **Limitations of current 2D pose detectors.** State-of-the-art pose estimators [45] produce noisy and incorrect results in frames with challenging motions, common in sports video. Here are examples from figure skating, tennis, gymnastics, and diving where pose estimates are incorrect.

We propose **Video Pose Distillation** (VPD), a weakly-supervised technique in which a *student* network learns to extract robust pose features from RGB video frames in a new video domain (a single sport). VPD is designed such that, whenever pose is reliable, the features match the output of a pretrained *teacher* pose detector. Our strategy retains the best of both pose and end-to-end worlds. First, like directly supervised end-to-end methods, our student can exploit the rich visual patterns present in the raw frames, including but not limited to the athlete's pose, and continue to operate when pose estimation is unsuccessful. Second, by constraining our descriptors to agree with the pose estimator whenever high-confidence pose is available, we avoid the pitfall of overfitting to visual patterns unrelated to the athlete's action. And third, weak pose supervision allows us to enforce an additional constraint: we require that the student predicts not only instantaneous pose but also its temporal derivative. This encourages our features to pick up on visual similarities over time (e.g., an athlete progress-

978-1-6654-2812-5/21/$31.00 ©2021 IEEE
DOI 10.1109/ICCV48922.2021.00912

ing from pose to pose). When we train the student with weak-supervision over a corpus of unlabeled sports video, the student learns to 'fill-in the gaps' left by the noisy pose teacher. Together, these properties lead to a student network whose features outperform the teacher's pose output when used in downstream applications.

VPD features improve performance on few-shot, fine-grained action recognition, retrieval, and detection tasks in the target sport domain, without requiring additional ground-truth action or pose labels. We demonstrate the benefits of VPD on four diverse sports video datasets with *fine-grained* action labels: diving [27], floor exercises [40], tennis [58], and a new dataset for figure skating. In a few-shot — limited supervision — setting, action recognition models trained with distilled VPD features can significantly outperform models trained directly on features from the teacher as well as baselines from prior skeleton-based and end-to-end learning work. For instance, when restricted to between 8 and 64 training examples per class from diving and floor exercises, the two datasets that are most challenging for pose, VPD features improve fine-grained classification accuracy by 6.8 to 22.8% and by 5.0 to 10.5%, respectively, over the next best method(s). Even when labels are plentiful, VPD remains competitive, achieving superior accuracy on three of the four test datasets. To summarize, VPD surpasses its teacher in situations where leveraging pose is crucial (e.g., few-shot) and is also competitive when end-to-end methods dominate (e.g., unreliable pose and the high-data / full supervision setting). Finally, we show applications of VPD features to fine-grained action retrieval and few-shot temporal detection tasks.

This paper makes the following contributions:

1. A weakly-supervised method, VPD, to adapt pose features to new video domains, which significantly improves performance on downstream tasks like action recognition, retrieval, and detection in scenarios where 2D pose estimation is unreliable.

2. State-of-the-art accuracy in few-shot, fine-grained action understanding tasks using VPD features, for a variety of sports. On action recognition, VPD features perform well with as few as 8 examples per class and remain competitive or state-of-the-art even as the training data is increased.

3. A new dataset (figure skating) and extensions to three datasets of real-world sports video, to include tracking of the performers, in order to facilitate future research on fine-grained sports action understanding.

2. Related Work

Pose representations provide a powerful abstraction for human action understanding. Despite significant progress in 2D and 3D pose estimation [36, 37, 45], downstream algorithms that depend on pose continue to suffer from unreliable estimates in sports video. With few labels available, for tasks such as fine-grained action recognition, models must learn both the actions and to cope with noisy inputs.

VIPE [44] and CV-MIM [61] show that learned pose embeddings, which factor-out camera view and forgo explicit 3D pose estimation, can be useful; they are trained on out-of-domain 3D pose data to embed 2D pose inputs and are effective when 2D pose is reliable. VPD extends these works by using distillation to replace the unreliable 2D pose estimation step with a model that embeds directly from pixels to pose-embedding. [22, 37, 59] learn human motion from video but produce 3D pose rather than embeddings.

Video action recognition is dominated by end-to-end models [3, 7, 13, 28, 43, 48, 51, 62], which are often evaluated on diverse but coarse-grained classification tasks (e.g., 'golf', 'tennis', etc.) [23, 25, 34, 42, 60]. Fine-grained action recognition in sports is a recent development [27, 40]. Besides being necessary for sports video analysis, fine-grained classification within a single sport is interesting because it avoids many contextual biases in coarse-grained tasks [9, 27, 52]. [2, 11, 16, 50] are also fine-grained datasets, but differ from body-centric actions in sports.

Pose or skeleton-based methods [10, 31, 57] appear to be a good fit for action recognition in human-centric sports. They depend on reliable 2D or 3D pose, which exists in datasets captured in controlled settings [30, 39] but not for public sports video, where no ground-truth is available and automatic detectors often perform poorly (e.g., [27, 40]).

VPD improves upon pose-based and end-to-end methods in human-centric sports datasets, especially when pose is not reliable. Like VIPE [44], VPD produces effective pose features, to the extent that comparatively simple downstream models such as nearest neighbor search [44] or a generic BiGRU [15] network can compete with the state-of-the-art in action recognition — in both few-shot and high-data regimes. To show this, we compare against several recent action recognition methods [31, 43] in Section 4.1.

VPD features can be used for any tasks where pretrained pose features may be helpful, such as action retrieval and temporally fine-grained detection (e.g., identifying tennis racket swings at 200 ms granularity). The latter is interesting because prior baselines [12, 21] focus on more general categories than human-centric action within a single sport and few papers [1, 56] address the few-shot setting.

Few-shot action recognition literature follows a number of paradigms, including meta-learning, metric learning, and data-augmentation approaches [1, 6, 26, 33]. These works focus on coarse-grained datasets [12, 23, 25, 42], adopt various protocols that partition the dataset into seen/unseen classes and/or perform a reduced N-way, K-shot classification (e.g., 5-way, 1 or 5 shot). VPD differs in that it is com-

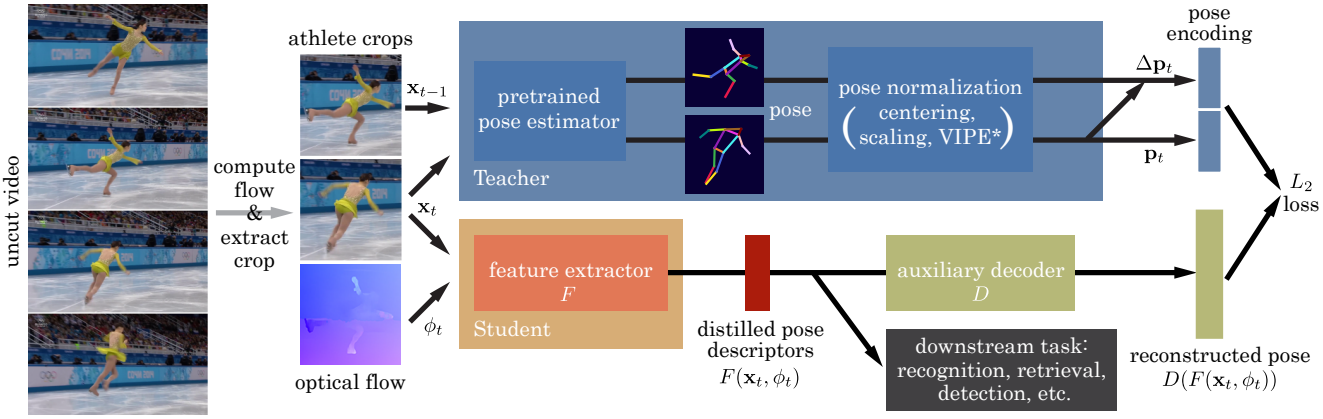

Figure 2: **Method overview.** VPD has two data pathways: a teacher to generate supervision and a student that learns to embed pose and motion in the target (sport) domain. When training on a frame t, the teacher applies an off-the-shelf 2D pose estimator, followed by a pose normalization step, to obtain weak pose features: $\mathbf{p}_t$ and $\Delta\mathbf{p}_t$. The student pathway receives the localized RGB $\mathbf{x}_t$ and optical flow ϕ_t, and computes a descriptor $F(\mathbf{x}_t, \phi_t) \in \mathbb{R}^d$, from which the fully connected network, D, regresses $(\mathbf{p}_t, \Delta\mathbf{p}_t)$. After training, only F is retained to extract embeddings on the full test dataset.

pletely agnostic to action labels when training features and does not require a particular architecture for downstream tasks such as action recognition. In contrast to 'few-shot' learning that seeks to generalize to unseen classes, *we evaluate on the standard classification task, with all classes known, but restricted to only k-examples per class at training time.* Our evaluation is similar to [41, 61], which perform action and image recognition with limited supervision, and, like [41, 61], we test at different levels of supervision.

Self-supervision/distillation. VPD relies on only machine-generated pose annotations for weak-supervision and distillation. VPD is similar to [55] in that the main goal of distillation is to improve the robustness and accuracy of the student rather than improve model efficiency. Most self-supervision work focuses on pretraining and joint-training scenarios, where self-supervised losses are secondary to the end-task loss, and subsequent or concurrent fine-tuning is necessary to obtain competitive results [8, 17, 20, 24, 29]. By contrast, our VPD student is fixed after distillation.

3. Video Pose Distillation

Our strategy is to distill inaccurate pose estimates from an existing, off-the-shelf pose detector — the *teacher* —, trained on generic pose datasets, into a — *student* — network that is specialized to generate robust pose descriptors for videos in a specific target sport domain (Figure 2). The student (Section 3.2) takes RGB pixels and optical flow, cropped around the athlete, as input. It produces a descriptor, from which we regress the athlete's pose as emitted by the teacher (Section 3.1). We run this distillation process over a large, *uncut and unlabeled* corpus of target domain videos (Section 3.3), using the sparse set of high-confidence teacher outputs as weak supervision for the student.

Since the teacher is already trained, VPD requires no new pose annotations in the target video domain. Likewise, no downstream application-specific labels (e.g., action labels for recognition) are needed to learn pose features. VPD does, however, require that the athlete be identified in each input frame, so we assume that an approximate bounding box for the athlete is provided in each frame as part of the dataset. Refer to Section 5 for discussion and limitations.

3.1. Teacher Network

To stress that VPD is a general approach that can be applied to different teacher models, we propose two teacher variants of VPD. The first uses an off-the-shelf pose estimator [45] to estimate 2D joint positions from $\mathbf{x}_t$, the RGB pixels of the t-th frame. We normalize the 2D joint positions by rescaling and centering as in [44], and we collect the joint coordinates into a vector $\mathbf{p}_t \in \mathbb{R}^d$. We refer to this as 2D-VPD since the teacher generates 2D joint positions.

Our second teacher variant further processes the 2D joint positions into a *view-invariant* pose descriptor, emitted as $\mathbf{p}_t$. Our implementation uses VIPE* to generate this descriptor. VIPE* is a reimplementation of concepts from Pr-VIPE [44] that is extended to train on additional synthetic 3D pose data [32, 38, 63] for better generalization. We refer to this variation as VI-VPD since the teacher generates a view-invariant pose representation. (See supplemental for details about VIPE* and its quality compared to Pr-VIPE.)

3.2. Student Feature Extractor

Since understanding an athlete's motion, not just their current pose, is a key aspect of many sports analysis tasks, we design a student feature extractor that encodes information about both the athlete's current pose $\mathbf{p}_t$ and the rate of change in pose $\Delta\mathbf{p}_t := \mathbf{p}_t - \mathbf{p}_{t-1}$.

The student is a neural network F that consumes a color video frame $\mathbf{x}_t \in \mathbb{R}^{3hw}$, cropped around the athlete, along with its optical flow $\phi_t \in \mathbb{R}^{2hw}$, from the previous frame. h and w are the crop's spatial dimensions, and t denotes the frame index. The student produces a descriptor $F(\mathbf{x}_t, \phi_t) \in \mathbb{R}^d$, with the same dimension d as the teacher's output. We implement F as a standard ResNet-34 [18] with 5 input channels, and we resize the input crops to 128×128.

During distillation, the features emitted by F are passed through an auxiliary decoder D, which predicts *both* the current pose $\mathbf{p}_t$ and the temporal derivative $\Delta\mathbf{p}_t$. Exploiting the temporal aspect of video, $\Delta\mathbf{p}_t$ provides an additional supervision signal that forces our descriptor to capture motion in addition to the current pose. D is implemented as a fully-connected network, and we train the combined student pathway $D \circ F$ using the following objective:

$$\underset{F,D}{\text{minimize}} \sum_{t=1}^{N} \left\| D\left(F\left(\mathbf{x}_t, \phi_t\right)\right) - \begin{bmatrix} \mathbf{p}_t \\ \Delta\mathbf{p}_t \end{bmatrix} \right\|_2^2 \qquad (1)$$

Since only F is needed to produce descriptors during inference, we discard D at the end of training.

Unlike its teacher, which was trained to recognize a general distribution of poses and human appearances, the student F *specializes* to frames and optical flow in the new target domain (e.g., players in tennis courts). Specialization via distillation allows F to focus on patterns present in the sports data that explain pose. We do not expect, nor do downstream tasks require, that F encode poses or people not seen in the target domain (e.g., sitting on a bench, ballet dancers), although they may be part of the teacher's training distribution. Experiments in Section 4 show that our pose descriptors, $F(\mathbf{x}_t, \phi_t)$, improve accuracy on several applications, including few-shot, fine-grained action recognition.

3.3. Training Data Selection and Augmentation

Data selection. The teacher's output may be noisy due to challenges such as motion blur and occlusion or because of domain shift between our target videos and the data that the teacher was trained on. To improve the student's ability to learn and to discourage memorization of the teacher's noise, we exclude frames with low pose confidence scores (specifically, *mean estimated joint score*) from the teacher's weak-supervision set. By default, the threshold is 0.5, although 0.7 is used for tennis. Tuning this threshold has an effect on the quality of the distilled features (see supplemental for details). We also withhold a fixed fraction of frames (20%) uniformly at random as a validation set for the student.

Data augmentation. We apply standard image augmentations techniques such as random resizing and cropping; horizontal flipping; and color and noise jitter, when training the student F. To ensure that left-right body orientations are preserved when horizontally augmenting $\mathbf{x}_t$ and ϕ_t, we also must flip the teacher's output $\mathbf{p}_t$. For 2D joint positions and 2D-VPD, this is straightforward. To flip VIPE* (itself a chiral pose embedding) used to train VI-VPD, we must flip the 2D pose inputs to VIPE* and then re-embed them.

4. Results

We evaluate the features produced by VPD on four fine-grained sports datasets that exhibit a wide range of motions.

Figure skating consists of 371 singles mens' and womens' short program performances from the Winter Olympics (2010-18) and World Championships (2017-19), totalling 17 video hours. In the classification task, FSJump6, there are six jump types defined by the ISU [19]. All videos from 2018 (134 routines, 520 jumps) are held out for testing. The remaining jumps are split 743/183 for training/validation.

Tennis consists of nine singles matches from two tournaments (Wimbledon and US Open), with swings annotated at the frame of ball contact [58]. There are seven swing classes in Tennis7. The training/validation sets contain 4,592/1,142 examples from five matches and the test set contains 2,509 from the remaining four matches. Split by match video, this dataset is challenging due to the limited diversity in clothing and unique individuals (10 professional players).

Floor exercise. We use the womens' floor exercise event (FX35) of the FineGym99 dataset [40], containing 1,214 routines (34 hours). There are 35 classes and 7,634 actions.

Diving48 [27] contains 16,997 annotated instances of 48 dive sequences, defined by FINA [14]. We evaluate on the corrected V2 labels released by the authors and retrain the existing state-of-the-art method, GSM [43], for comparison.

All four datasets contain frames in which pose is not well estimated or uncertain, though their distribution varies (see supplemental for details). As mentioned beforehand, pose estimates are typically worse in frames with fast motion, due to motion blur and atypical, athletic poses such as flips or dives; see Figure 1 for examples. A common challenge across these datasets, the fast-motion frames are often necessary for discriminating the fine-grained actions of interest.

We assume the subject of the action is identified and tracked. With multiple humans in the frame, fast-moving athletes in challenging poses are often missed otherwise: i.e., detected at lower confidence than static audience members or judges, or not detected at all. *For fair comparison, we boost the baselines by providing them the same inputs as our method, which improves their results significantly.*

4.1. Fine-Grained Action Recognition

Fine-grained action recognition tests VPD's ability to capture precise details about an athlete's pose and motion. We consider both the few-shot setting, where only a limited number of action examples are provided, and the traditional

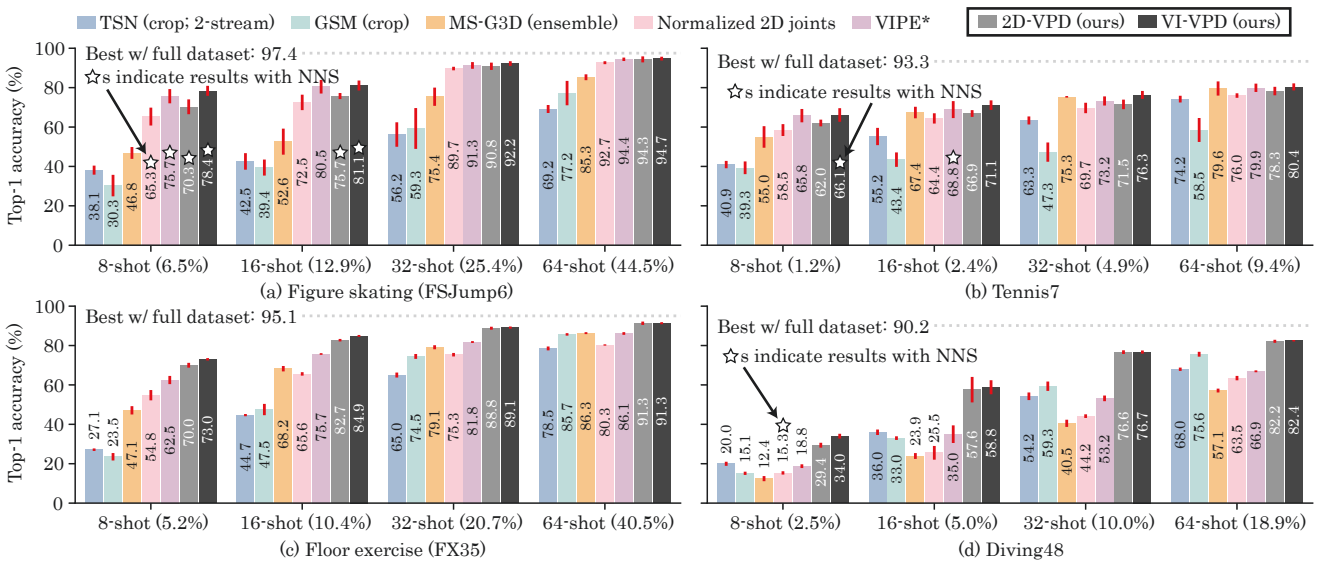

Figure 3: **Accuracy on few-shot fine-grained action recognition.** Percentages give the fraction of the full training set. State-of-the-art accuracies using the full dataset as supervision are indicated as a dashed line for reference (see Table 1). Pose-based baselines (MS-G3D [31], 2D joints, and VIPE*) surpass end-to-end baselines (GSM [43] and TSN [51]) in few-shot settings on every dataset except Diving48, demonstrating both the importance of pose when labels are scarce and the challenge when pose is unreliable. VI-VPD significantly outperforms the baselines and prior methods on FX35 and Diving48; accuracy on FSJump6 and Tennis7 also improves slightly, but remains similar to VIPE*. *Starred results above use nearest-neighbor search (NNS) instead of the BiGRU architecture (NNS performed better in these cases; see supplemental for full results).*

full supervision setting, where all of the action examples in the training set are available.

Our VPD features are distilled over the training videos in the sports corpus, uncut and without labels. To extract features on the test set, we use the fixed VPD student F. VI-VPD and 2D-VPD features maintain the same dimensions d, of their teachers: $d = 64$ for VIPE* and $d = 26$ for normalized 2D joints. For Diving48, VIPE* has $d = 128$ because we also extract pose embeddings on the vertically flipped poses and concatenate them. This data augmentation is beneficial for VIPE* due to the often inverted nature of diving poses, which are less well represented in the out-of-domain 3D pose datasets that VIPE* is trained on.

Action recognition model. To use VPD for action recognition, we first represent each action as a sequence of pose features. We then classify actions using a bidirectional Gated Recurrent Unit network (BiGRU) [15] trained atop the (fixed) features produced by the student F. Since our features are chiral and many actions can be performed with either left-right orientation, we embed both the regular and horizontally flipped frames with the student. See supplemental for implementation details.

Prior pose embedding work has explored using sequence alignment followed by nearest-neighbor retrieval [44]. We also tested a nearest-neighbor search (NNS) approach that uses dynamic time warping to compute a matching cost between sequences of pose features. For NNS, each test example is searched against all the training examples, and the label of the best aligned match is predicted. The BiGRU is superior in most settings, though NNS can be effective in few-shot situations, and we indicate when this is the case.

Baselines. We compare our distilled 2D-VPD and VI-VPD features against several baselines.

1. The *features from the teacher:* VIPE* or the normalized 2D joint positions, using the same downstream action recognition models and data augmentations.

2. *Skeleton-based:* a MS-G3D ensemble [31] and ST-GCN [57]. Both baselines receive the same tracked 2D poses used to supervise VPD.

3. *End-to-end:* GSM [43], TSN [51], and TRNms [62] (multiscale). We test with both the cropped athletes and the full frame (w/o cropping) as inputs, and we find that cropping significantly improves accuracy in both the few-shot setting on all four datasets, and the full supervision setting on all datasets except for Diving48. When applicable, combined results with RGB and optical flow models are indicated as 2-stream.

4.1.1 Few-shot and limited supervision setting

Experiment protocol. Each model is presented k examples of each action class but may utilize unlabeled data

Dataset (Top-1 acc)	FSJump6	Tennis7	FX35	Diving48
Random	16.7	14.3	2.9	2.1
Top class	33.7	46.7	7.5	8.3
End-to-end				
†TRNms (2-stream) [40, 62]			84.9	
†TimeSformer-L [3]				81.0
TSN [51] (w/o crop)	57.9	-	83.2	82.3
TSN (crop)	81.2	87.8	88.5	83.6
TSN (crop; 2-stream)	82.7	90.9	90.4	83.6
TRNms [62] (w/o crop)	68.7	-	81.5	80.5
TRNms (crop)	77.7	55.5	87.1	81.8
TRNms (crop; 2-stream)	84.0	76.3	87.3	81.5
GSM [43] (w/o crop)	42.1	-	90.3	**90.2**
GSM (crop)	90.6	67.1	93.6	88.7
Skeleton / pose-based (w/ tracked 2D poses)				
†ST-GCN (w/o tracking) [40, 57]			40.1	
ST-GCN [57]	88.7	88.4	80.3	64.8
MS-G3D (ensemble) [31]	91.7	91.0	92.1	80.2
Pose features (w/ BiGRU)				
Normalized 2D joints	95.5	90.9	86.9	75.7
(Ours) 2D-VPD	97.0	92.6	94.5	86.4
VIPE⋆	96.8	91.8	90.8	78.6
(Ours) VI-VPD	**97.4**	**93.3**	94.6	88.6
(Ours) Concat-VPD	96.2	93.2	**95.1**	88.7

Table 1: **Accuracy on fine-grained action recognition with all of the training data.** Top results overall are **bolded** and per method category are underlined. † indicates best results from prior work. VI-VPD achieves SOTA accuracy on FSJump6, Tennis7, and FX35, even when baselines are improved with tracking and cropped inputs. On Diving48, VI-VPD trails end-to-end GSM (w/o crop) by 1.6%. VI-VPD and 2D-VPD features can both be competitive; concatenating them (Concat-VPD) may improve accuracy slightly.

or knowledge from other datasets as pre-training. For example, skeleton-based methods rely on 2D pose detection; VIPE⋆ leverages out-of-domain 3D pose data; and VPD features are distilled on the uncut, unlabeled training videos. This experimental setup mirrors real-world situations where few labels are present but unlabeled and out-of-domain data are plentiful. Our evaluation metric is top-1 accuracy on the full test set. To control for variation in the training examples selected for each few-shot experiment, we run each algorithm on five randomly sampled and fixed subsets of the data, for each k, and report the mean accuracy.

Results. Figure 3 compares 2D-VPD and VI-VPD features to their teachers (and other baselines). On FSJump6 and Tennis7, VI-VPD provides a slight improvement over its state-of-the-art teacher, VIPE⋆, with accuracies within a few percent. FX35 shows a large improvement and VI-VPD increases accuracy by up to 10.5% over VIPE⋆ at $k \leq 32$ and 5% over the MS-G3D ensemble at $k = 64$. Likewise, on Diving48, where end-to-end GSM and 2-stream TSN are otherwise better than the non-VPD pose-based methods, VI-VPD improves accuracy by 6.8 to 22.8%. Our results on FX35 and Diving48 suggest that VI-VPD helps to transfer the benefits of pose to datasets where it is most unreliable.

While view-invariant (VI) features generally perform better than their 2D analogues, the difference in accuracy between VI-VPD and 2D-VPD is more noticeable in sports with diverse camera angles (such as figure skating and floor exercise) and at small k, where the action recognition model can only observe a few views during training.

4.1.2 Traditional, full training set setting

VPD features are competitive even in the high-data regime (Table 1). On all four datasets, both VI-VPD and 2D-VPD significantly improve accuracy over their teachers. VI-VPD also achieves state-of-the-art accuracy on the FSJump6 (0.6% over VIPE⋆), Tennis7 (1.5% over VIPE⋆), and FX35 (1.0% over GSM, with cropped inputs) datasets.

Diving48 is especially challenging for pose-based methods, and VI-VPD performs worse than GSM, without cropping, by 1.6%. GSM, with cropping, is also worse by 1.5%, possibly due to errors and limitations of our tracking. VI-VPD does, however, perform significantly better than the top pose-based baseline (8.4% over MS-G3D, ensemble).

Our results demonstrate that VPD's success is not limited to few-shot regimes. However, because many methods in Table 1 can produce high accuracies, at or above 90%, when given ample data, we view improving label efficiency as a more important goal for VPD and future work.

4.1.3 Ablations and additional experiments

We highlight two important ablations of VPD to understand the source of VPD's improvements: (1) analyzing parts of the distillation method and (2) distilling with only the action segments of the video. We also consider (3) an unlabeled setting where VPD is distilled over the entire video corpus. Please refer to supplemental for additional experiments.

Analysis of the distillation method. Table 2(a) shows the increase in accuracy on action recognition for ablated 2D-VPD and VI-VPD features when we distill without flow input ϕ_t and without motion prediction[1]. The incremental improvements are typically most pronounced in the few-shot setting, on the FX35 and Diving48 datasets, where VPD produces the largest benefits (see Section 4.1.1).

[1]The student mimics the teacher's $\mathbf{p}_t$ output directly, without the auxiliary decoder D and $\Delta\mathbf{p}_t$ in the training loss.

Dataset	FSJump6		Tennis7		FX35		Diving48	
Input features \ Amount of training data	Full	16-shot	Full	16	Full	16	Full	16
(a) Normalized 2D joints (teacher)	95.5	72.5	90.9	64.3	86.9	65.6	75.7	25.5
distilled w/o motion; RGB	96.1	73.2	90.9	66.5	92.0	76.3	85.3	52.8
distilled w/o motion; RGB & flow	95.8	74.6	91.7	67.0	91.6	76.6	85.6	53.0
2D-VPD: distilled w/ motion; RGB & flow	97.0	74.4	92.6	66.9	94.5	82.7	86.4	57.6
VIPE* (teacher)	96.8	80.5	91.8	67.0	90.8	75.7	78.6	35.0
distilled w/o motion; RGB	97.1	**81.3**	92.1	67.6	93.5	83.4	86.5	54.9
distilled w/o motion; RGB & flow	97.3	79.3	91.7	69.7	92.9	83.2	85.9	53.7
VI-VPD: distilled w/ motion; RGB & flow	**97.4**	80.2	**93.3**	**71.1**	**94.6**	**84.9**	**88.6**	**58.8**
(b) VI-VPD (distilled on action video only)	96.3	79.4	92.4	69.1	94.1	84.3	-	-
(c) VI-VPD (distilled w/ the entire video corpus)	97.2	81.9	93.8	72.6	94.5	84.9	88.4	59.6

Table 2: **Action recognition experiments.** Top-1 accuracy in the full training set and 16-shot scenarios with (a) ablations to the distillation methodology, (b) when only the action parts of the dataset are used for distillation, and (c) when VI-VPD features are distilled over the entire video corpus (including the testing videos, without labels). Results are with the BiGRU.

With VIPE* as the teacher, distillation alone from RGB can have a large effect (2.7% and 7.7%, at full and 16-shot settings on FX35; 7.9% and 19.9% on Diving48). Adding flow in addition to RGB, without motion, gives mixed results. Finally, adding motion prediction and decoder D, further improves results (1.1% and 1.5% on FX35, at full and 16-shot; 2.1% and 3.9% on Diving48). The effect of distilling motion on FSJump6 and Tennis7 is mixed at the 16-shot setting, though the full setting shows improvement.

2D-VPD can be seen as an ablation of view-invariance (VIPE*) and shows a similar pattern when further ablated.

Training VPD on action parts of video only. Fine-grained action classes represent less than 7%, 8%, and 28% of the video in FSJump6, FX35, and Tennis7. We evaluate whether distillation of VI-VPD over uncut video improves generalization on action recognition, by distilling VI-VPD features with *only the action parts* of the training videos.

The results are summarized in Table 2(b) and show that distilling with only the action video performs worse on our datasets. This is promising because (1) uncut performances are much easier to obtain than performances with actions detected, and (2) in the low-supervision setting, VI-VPD improves accuracy even if actions have not been detected in the rest of the training corpus. This also suggests that distilling over more video improves the quality of the features.

Distillation with the entire video corpus. An unlabeled corpus is often the starting point when building real-world applications with videos in a new domain (e.g., [58]). Because VPD is supervised only by machine-generated pose estimates from unlabeled video, VPD features can be distilled over all of the video available, not just the training data.[2] Table 2(c) shows results when VI-VPD is distilled jointly with both the training and testing videos, *uncut and without labels*. The improvement, if any, is minor on all four datasets ($\leq$1.5%, attained on Tennis7 at 16-shot) and demonstrates that VI-VPD, distilled over a large dataset, is able to generalize without seeing the test videos.

[2]This setting is similar to [46, 47], which propose self-supervision to align the training and testing distributions in situations with large domain shift.

4.2. Action Retrieval

Action retrieval measures how well VPD features can be used to search for similar unlabeled actions. Here, the VPD features are distilled on the entire unlabeled corpus.

Experiment protocol. Given a query action, represented as a sequence of pose features, we rank all other actions in the corpus using the L_2 distance between pose features and dynamic time warping to compute an alignment score. A result is considered relevant if it has the same fine-grained action label as the query, and we assess relevance by the precision at k results, averaged across all the queries.

Results. At all cut-offs in Table 3 and in all four datasets, VPD features outperform their teachers. Sizeable improvements are seen on FX35 and Diving48. View-invariance does not always result in the highest precision if the number of camera angles is limited (e.g., Tennis7 and Diving48), though it may be helpful for retrieving more diverse results.

4.3. Pose Features for Few-Shot Action Detection

Detection of fine-grained actions, at fine temporal granularity and with few labels, enables tasks such as few-shot recognition and retrieval. We evaluate VPD features on the figure skating and tennis datasets, to temporally localize the jumps and the swings, respectively. The average jump is 1.6 seconds in length ($\approx$40 frames), while a swing is defined to be the 200 ms around the frame of ball contact ($\approx$5 frames).

Experiment protocol. *We follow the same video-level train/test splits as FSJump6 and Tennis7, and distill features*

Dataset	FSJump6			Tennis7			FX35			Diving48		
k	1	10	50	1	10	50	1	10	50	1	10	50
Normalized 2D joints	91.8	84.8	73.8	91.8	88.1	82.1	71.6	57.4	39.0	34.5	22.1	14.6
(Ours) 2D-VPD	92.5	86.4	76.3	**93.1**	90.0	84.6	79.7	66.8	47.5	**64.4**	**43.8**	**27.9**
VIPE⋆	92.9	85.1	75.7	92.4	90.0	85.9	72.2	60.1	46.6	36.1	24.1	15.1
(Ours) VI-VPD	**93.6**	**86.8**	**78.0**	92.8	**90.6**	**86.3**	**80.8**	**68.6**	**52.4**	60.9	40.9	25.4

Table 3: **Action retrieval: Precision@k results (%) ranked by alignment score with dynamic time warping.** VPD leads to more relevant results on all four datasets. Gains on FSJump6 and Tennis7 are modest, while the large improvements on FX35 and Diving48 suggest that VPD features are superior in cases when pose estimates are the most unreliable.

Temporal IoU	0.3	0.4	0.5	0.6	0.7
Figure skating jumps (trained on five routines)					
Pretrained R3D [49]	39.5	30.0	23.1	15.0	9.0
Normalized 2D joints	80.6	70.0	53.5	40.2	24.6
(Ours) 2D-VPD	85.7	77.8	**61.5**	47.6	25.8
VIPE⋆	84.5	76.8	59.3	45.3	26.7
(Ours) VI-VPD	**86.1**	**78.6**	60.7	**47.9**	**28.7**
Tennis swings at 200 ms (trained on five points)					
Pretrained R3D [49]	41.3	37.8	29.9	15.8	7.6
Normalized 2D joints	59.7	58.2	43.7	24.6	10.3
(Ours) 2D-VPD	67.4	66.5	54.0	28.4	13.1
VIPE⋆	67.4	65.8	51.2	28.9	12.3
(Ours) VI-VPD	**73.5**	**72.6**	**58.6**	**32.9**	**13.8**

Table 4: **Few-shot action detection: Average precision (AP) at various levels of temporal IoU.** VI-VPD features improve AP over VIPE⋆ and the other baselines.

on the training videos only. As a simple baseline method, we train a BiGRU that outputs per-frame predictions, which are merged to produce predicted action intervals (see supplemental for details). The BiGRU is trained on ground-truth temporal labels from five routines (figure skating) and five points (tennis). For more consistent results, we perform five-fold cross-validation and ensemble the per-frame predictions. In Table 4, we report average precision (AP) at various levels of temporal intersection over union (tIoU).

Results. VPD improves AP on both tasks. The short duration of tennis swings means that noise in per-frame pose estimates has a large impact, and VI-VPD improves AP at every tIoU threshold (up to 7.4 over VIPE⋆ at tIoU $= 0.5$).

5. Limitations and Discussion

Subject tracking is needed for VPD to ensure that the pose is of the correct person. Real-world sports video often contains many people, such as audience and judges, in addition to the subject. The tracking annotations in the datasets in Section 4.1 are computed automatically using off-the-shelf models and heuristics (see supplemental for details). This is possible because athletes are salient in appearance, space, and time — sports video is a natural application for work on tracking [4, 53] and detecting salient regions [5]. We observe that the difference in accuracy between the tracked and non-tracked inputs on other prior methods such as [43, 51, 57] can be staggering (48% on FSJump6 for GSM [43] and 40% on FX35 for ST-GCN [57]; see Table 1).

To evaluate the quality of our pose features, we focused on motion by a single athlete or synchronized athletes (contained in Diving48). Tasks and actions involving many people require a more sophisticated downstream model that can handle multiple descriptors or poses per frame.

Future work. First, the 2D pose estimates used to supervise VPD are inherently ambiguous with respect to camera view, and additional information such as depth or a behavioral prior could help alleviate this ambiguity. Other weak supervision sources, in addition to motion and VIPE, may also help. Second, our distillation process is offline; supporting online training, similar to [35, 47], at the pose feature extraction stage could be beneficial in time-evolving datasets. Distillation for explicit 2D or 3D pose estimation is another possibility. Although VPD features can improve accuracy with limited data, performance on few-shot and semi-supervised tasks still has much room to improve, and we hope that future work continues to explore these topics.

6. Conclusion

Pose features are useful for studying human-centric action in novel sports video datasets. However, such datasets are often challenging for off-the-shelf models. Our method, VPD, improves the reliability of pose features in difficult and label-poor settings, by distilling knowledge from existing pose estimators. VPD learns features that improve accuracy on both traditional and few-shot action understanding tasks in the target (sport) domain. We believe that our distillation-based method is a useful paradigm for addressing challenges faced by applications in new video domains.

Acknowledgements. This work is supported by the National Science Foundation (NSF) under III-1908727 and Adobe Research. We also thank the anonymous reviewers.

References

[1] Rami Ben-Ari, Mor Shpigel Nacson, Ophir Azulai, Udi Barzelay, and Daniel Rotman. TAEN: Temporal Aware Embedding Network for Few-Shot Action Recognition. In *CVPR Workshops*, 2021. 2

[2] Yizhak Ben-Shabat, Xin Yu, Fatemeh Sadat Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The IKEA ASM Dataset: Understanding People Assembling Furniture through Actions, Objects and Pose, 2020. arXiv:2007.00394. 2

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? In *ICML*, 2021. 2, 6

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple Online and Realtime Tracking. In *ICIP*, 2016. 8

[5] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where Should Saliency Models Look Next? In *ECCV*, 2016. 8

[6] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-Shot Video Classification via Temporal Alignment. In *CVPR*, 2020. 2

[7] J. Carreira and A. Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017. 1, 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 2020. 3

[9] Jinwoo Choi, Chen Gao, C. E. Joseph Messou, and Jia-Bin Huang. Why Can't I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition. In *NeurIPS*, 2019. 1, 2

[10] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. PoTion: Pose MoTion Representation for Action Recognition. In *CVPR*, 2018. 2

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *ECCV*, 2018. 2

[12] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *CVPR*, 2015. 2

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *ICCV*, 2019. 1, 2

[14] FINA. Fédération Internationale de Natation. 4

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*, chapter Sequence Modeling: Recurrent and Recursive Nets. MIT Press, 2016. http://www.deeplearningbook.org. 2, 5

[16] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *ICCV*, 2017. 2

[17] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised Co-Training for Video Representation Learning. In *NeurIPS*, 2020. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2015. 4

[19] ISU. International Skating Union. 4

[20] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video Representation Learning by Recognizing Temporal Transformations. In *ECCV*, 2020. 3

[21] Yu-Gang Jiang, Jingen Liu, Amir Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. THUMOS Challenge: Action Recognition with a Large Number of Classes, 2014. 2

[22] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D Human Dynamics from Video. In *CVPR*, 2019. 2

[23] 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. arXiv:1705.06950. 2

[24] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles. In *AAAI*, 2019. 3

[25] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A Large Video Database for Human Motion Recognition. In *ICCV*, 2011. 2

[26] Sai Kumar Dwivedi, Vikram Gupta, Rahul Mitra, Shuaib Ahmed, and Arjun Jain. ProtoGAN: Towards Few Shot Learning for Action Recognition. In *ICCV Workshops*, 2019. 2

[27] Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: Towards Action Recognition without Representation Bias. In *ECCV*, 2018. 2, 4

[28] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal Shift Module for Efficient Video Understanding. In *ICCV*, 2019. 1, 2

[29] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. MS2L: Multi-Task Self-Supervised Learning for Skeleton Based Action Recognition. In *ACM Multimedia*, 2020. 3

[30] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. 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, 2020. 2

[31] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *CVPR*, 2020. 1, 2, 5, 6

[32] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of Motion Capture as Surface Shapes. In *ICCV*, 2019. 3

[33] Ashish Mishra, Vinay Kumar Verma, M. Shiva Krishna Reddy, Arulkumar Subramaniam, Piyush Rai, and Anurag Mittal. A Generative Approach to Zero-Shot and Few-Shot Action Recognition. In *WACV*, 2018. 2

[34] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):502–508, 2020. 2

[35] R. Mullapudi, S. Chen, K. Zhang, D. Ramanan, and K. Fatahalian. Online Model Distillation for Efficient Video Inference. In *ICCV*, 2019. 8

[36] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. In *ECCV*, 2018. 2

[37] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019. 2

[38] Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3DPeople: Modeling the Geometry of Dressed Humans. In *ICCV*, 2019. 3

[39] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *CVPR*, 2016. 2

[40] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. FineGym: A Hierarchical Video Dataset for Fine-Grained Action Understanding. In *CVPR*, 2020. 2, 4, 6

[41] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *NeurIPS*, 2020. 3

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, 2012. arXiv:1212.0402. 2

[43] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-Shift Networks for Video Action Recognition. In *CVPR*, 2020. 1, 2, 4, 5, 6, 8

[44] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-Invariant Probabilistic Embedding for Human Pose. In *ECCV*, 2020. 2, 3, 5

[45] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. In *CVPR*, 2019. 1, 2, 3

[46] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A. Efros. Unsupervised Domain Adaptation through Self-Supervision, 2019. arXiv:1909.11825. 7

[47] Yu Sun, Xiaolong Wang, Liu Zhuang, John Miller, Moritz Hardt, and Alexei A. Efros. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *ICML*, 2020. 7, 8

[48] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *CVPR*, 2018. 2

[49] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *CVPR*, 2018. 8

[50] TwentyBN. The 20BN-something-something Dataset V2. 2

[51] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Val Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV*, 2016. 1, 2, 5, 6, 8

[52] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context, 2021. arXiv:1912.07249. 1, 2

[53] Nicolai Wojke and Alex Bewley. Deep Cosine Metric Learning for Person Re-identification. In *WACV*, 2018. 8

[54] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[55] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-Training With Noisy Student Improves ImageNet Classification. In *CVPR*, 2020. 3

[56] Huijuan Xu, Bingyi Kang, Ximeng Sun, Jiashi Feng, Kate Saenko, and Trevor Darrell. Similarity R-C3D for Few-shot Temporal Activity Detection. *arXiv*, 2018. arXiv:1812.10000. 2

[57] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI*, 2018. 1, 2, 5, 6, 8

[58] Haotian Zhang, Cristobal Sciutto, Maneesh Agrawala, and Kayvon Fatahalian. Vid2Player: Controllable Video Sprites That Behave and Appear Like Professional Tennis Players. *ACM Transactions on Graphics*, 40(3), 2021. 2, 4, 7

[59] Jason Y. Zhang, Panna Felsen, Angjoo Kanazawa, and Jitendra Malik. Predicting 3D Human Dynamics from Video. In *ICCV*, 2019. 2

[60] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding. In *ICCV*, 2013. 2

[61] Long Zhao, Yuxiao Wang, Jiaping Zhao, Liangzhe Yuan, Jennifer J. Sun, Florian Schroff, Hartwig Adam, Xi Peng, Dimitris Metaxas, and Ting Liu. Learning View-Disentangled Human Pose Representation by Contrastive Cross-View Mutual Information Maximization. In *CVPR*, 2021. 2, 3

[62] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal Relational Reasoning in Videos. In *ECCV*, 2018. 1, 2, 5, 6

[63] Luyang Zhu, Konstantinos Rematas, Brian Curless, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Reconstructing NBA Players. In *ECCV*, 2020. 3