Boosting Semi-supervised Video Action Detection with Temporal Context

Donghyeon Kwon[1] Inho Kim[1] Suha Kwak[1,2]
Dept. of CSE, POSTECH[1] Graduate School of AI, POSTECH[2]

Abstract

This paper studies semi-supervised learning of video action detection (VAD), which assumes that only a small portion of training videos are labeled and the others remain unlabeled. The existing semi-supervised methods for VAD mainly focus on leveraging spatial context of unlabeled video, lacking its exploration of temporal context. To resolve this, we present a novel semi-supervised learning framework that effectively incorporates spatio-temporal context during training. We first introduce a new augmentation strategy called temporal cross-view augmentation to achieve robust representation across clips depicting the same action but not aligned on the time axis. We also propose a new context fusion method called global-local context fusion that effectively utilizes the spatio-temporal context of videos to enhances the features of each frame by incorporating those of other frames within a clip; this method aids in actively leveraging spatio-temporal context of video, leading to significant performance improvement. Our framework was evaluated on UCF101-24 and JHMDB-21, where it outperformed all existing methods in every evaluation setting.

1. Introduction

Video action detection (VAD), the task of localizing actors in space-time and classifying his/her action at once, plays crucial roles in a wide range of applications such as video surveillance [8, 36, 39], human-computer interaction [11, 38] and healthcare [12, 19, 44]. Recent advances in VAD have been attributed to supervised learning of deep neural networks [4, 9, 17, 18, 22, 33, 41, 42, 46, 53, 58, 68, 70] on large-scale datasets [14, 21, 57]. However, collecting such labeled videos demands manual frame-wise annotation, which is prohibitively costly and thus often leads to training data lacking in both class diversity and quantity. To resolve this issue, we study semi-supervised learning for VAD, in which we suppose that only a subset of training videos are assigned manual labels while the others remain unlabeled. Also, following the convention of the previous work [25, 72], we assume that each annotated video is assigned both a video-level action class label and a frame-level localization label.

At the core of semi-supervised learning lies the way of using unlabeled data effectively for training. To this end, semi-supervised learning methods for image understanding models have generally employed consistency regularization [10,24,40,54] and contrastive learning [1,28,29,73,74]. Specifically, consistency regularization forces a model to produce consistent predictions when it encounters different views of the same inputs, even if the inputs are perturbed or augmented, while contrastive learning encourages unlabeled data of the same pseudo label to be close to each other in an embedding space. Although these techniques have proven effective, they are not optimal for semi-supervised learning in video since they primarily focus on learning in an image domain and, when directly applied to a video, do not actively leverage temporal information.

For semi-supervised learning in video, it is crucial to incorporate spatio-temporal context effectively. However, prior arts for semi-supervised VAD [25, 72] have been lacking in its exploration of temporal context during training; they utilize spatial context only, applying consistency regularization between two differently augmented clips sharing the same timestamp.[1] This approach leads to inconsistent representations across different clips depicting the same action but not aligned temporally, where the action occurs at disparate time intervals or manifests in different ways, resulting in diminished performance.

To address this issue and incorporate the temporal context into the spatial one, we present a novel framework for semi-supervised VAD, which is illustrated in Fig. 1. Our framework introduces a new augmentation strategy, dubbed *temporal cross-view augmentation*, to achieve robust feature representation across clips characterizing the same action but not aligned in the temporal domain. The proposed strategy first samples two clips from the same video so that they have different timestamps but share a subset of frames, and then performs consistency regularization on the shared frames. Furthermore, we deploy contrastive learning to those clips to further utilize frames not shared between the clips. Along with consistency regularization applied to

[1]The term 'timestamp' refers to the start and end times of a video.

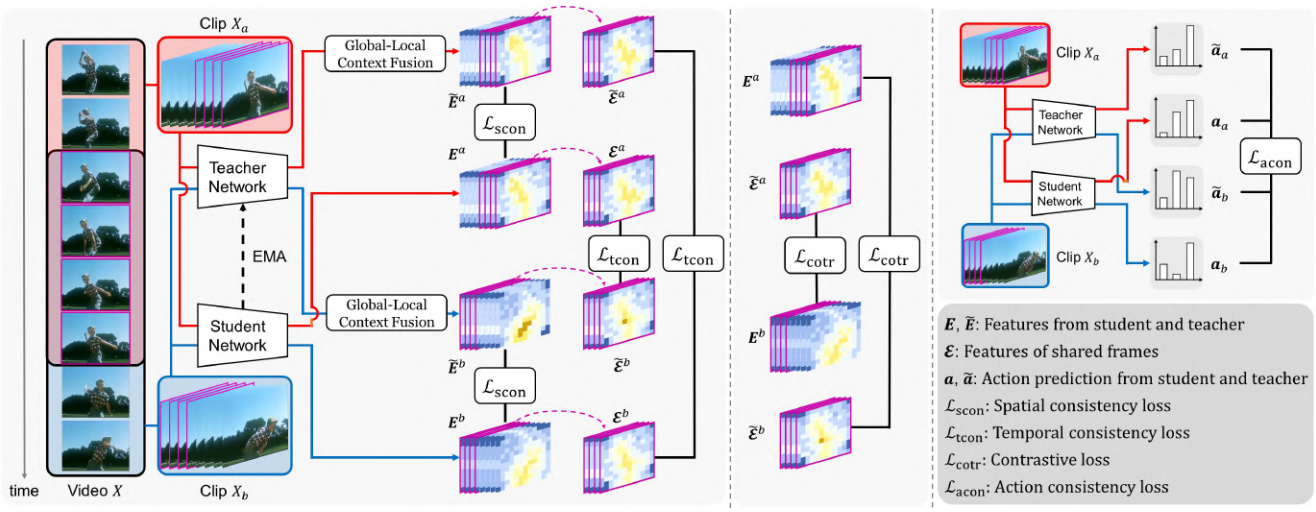

Figure 1. Our semi-supervised learning framework incorporates temporal cross-view augmentation and global-local context fusion. We first sample two clips X_a and X_b from the Video X so that they have different timestamps with shared frames. Each clip is fed into the student and teacher networks, *yielding their action predictions and pixel-wise feature embeddings at the same time, respectively.* Next, we apply global-local context fusion to the feature embeddings of the teacher and perform spatio-temporal consistency and contrastive learning with those of the teacher and raw feature embeddings from the student. Shared frames are highlighted with coloured boundaries in magenta.

two differently augmented versions of an unlabeled video that share the same timestamp in a frame-by-frame manner for the spatial domain, this strategy improves the model's performance thanks to its feature representation robust to spatio-temporal variations.

Moreover, we propose a new context fusion method, called *global-local context fusion*, that utilizes the spatio-temporal context of video for learning with unlabeled videos effectively. Based on the observation that different frames in a video exhibit distinct spatio-temporal information that complements each other [30, 37, 49, 61], our context fusion method enhances latent features of each frame by incorporating those of other frames within a clip. The basic idea is to extract the latent features of different frames within the same clip, combining them to capture a more global perspective, and fusing the global context with the features of each frame. In this fusion process, we aggregate and propagate information along the temporal axis in a time-ordered manner as the temporal evolution of a video provides key cues for understanding its content. This enables to integration of the global context with the local features of each frame.

Our method was evaluated on two public benchmarks, UCF101-24 [57] and JHMDB-21 [21], where it achieved the state-of-the-art performance on all the two datasets. In brief, our contribution is three-fold:

- We introduce a novel semi-supervised learning framework for video action detection with the proposed *temporal cross-view augmentation*, along with its proper learning strategy.
- We also propose an effective spatio-temporal context fusion method, *global-local context fusion*, enabling the model to enhance the latent features of each frame, resulting in a significant performance improvement.
- Our method achieved the state of the art on two public benchmarks, UCF101-24 and JHMDB-21, across all experimental settings.

2. Related Work

Video Action Detection. The task of video action detection is to localize actors in space-time and classify its action simultaneously. Significant developments have been made in video action detection in recent years [6, 13, 33, 34, 43, 46, 53, 62, 68, 70] due to advancements in deep neural networks. Early methods [13, 62] introduced a two-stream 2D model using an RGB frame and optical flow. [22] proposes a detector that takes a sequence of frames as input and outputs tubelets. [17] exploits 3D convolutional layers to capture motion characteristics in videos. [68] introduces gradually updating initial proposals during training. [56] utilizes a recurrent mechanism with ConvLSTM [66] to consider longer temporal information. Other works [14, 70] incorporate optical flow into a neural network. Most of the listed methods utilize a proposal-based approach [14, 17, 68, 70], leading to a complex two-step process, while [9] introduces a simple end-to-end approach based on capsule routing.

Semi-supervised Learning. Reducing the cost of labeling in machine learning through the use of semi-supervised

learning is an active area of research. Consistency regularization is widely studied for semi-supervised learning [2, 20, 54, 59, 67]. It improves model performance by encouraging the model to make consistent outputs from an input and its augmented version. Another approach, contrastive learning [1, 7, 15, 27, 28, 50], encourages unlabeled data with the same pseudo label to be close to each other in an embedding space, showing performance improvement. Some methods [51, 64] for semi-supervised video action recognition were also introduced. [51] mainly uses consistency learning between predictions from two differently sub-sampled clips, while [64] leverages temporal gradients as another input to further utilize temporal context.

Semi-supervised Video Action Detection. Recently, various methods in semi-supervised video action detection have been introduced. [25] is the first work in the literature that employs a consistency-based approach in a frame-by-frame manner. Building upon this method, [72] further investigates consistency in the video background region and proposes background-weakening to reduce false detection on the background. [52] introduces active learning strategy to leverage labeled data in a semi-supervised manner, while [69] utilizes pseudo labels with re-weighting strategy. However, the listed methods mainly focus on spatial consistency, emphasizing frame-by-frame consistency regularization, leaving room for effective utilization of temporal context for video action detection.

3. Proposed Method

On labeled videos, our framework is trained with supervised learning losses for action classification and space-time localization. For unlabeled videos, we use consistency regularization and contrastive learning in both spatial and temporal domains via our *temporal cross-view augmentation* and *global-local context fusion* strategies. Our framework comprises three key components: supervised learning with labeled data (Section 3.2), spatio-temporal semi-supervised learning with temporal cross-view augmentation (Section 3.3), and the global-local context fusion (Section 3.4).

3.1. Preliminaries

Let c denote the number of action classes. Given a video comprising n frames of $h \times w$ size, denoted as $X = (v_1, \dots, v_n)$, and associated with ground truth $Y = (\mathcal{Y}, \mathbf{y})$, where $\mathcal{Y} \in \{0,1\}^{n \times h \times w}$ and $\mathbf{y} \in \{1, 2, ..., c\}$ represent the pixel-wise localization map and action class label for individual frames of the video, our objective is to train a VAD model with both a small labeled video set $D_L = \{(X_i, Y_i)\}_{i=1}^{N_l}$ and a large collection of unlabeled videos, denoted as $D_U = \{X_j\}_{j=1}^{N_u}$, where N_l and N_u denotes the number of data in D_L and D_U, respectively.

3.2. Supervised Learning with Labeled Data

Our model is trained using labeled data with two losses: an action classification loss and a space-time localization loss. In line with previous work [9, 26, 72], we employ the spread loss [9] for action classification and a combination of the cross-entropy loss and the dice loss for space-time localization. For an input video X, let $\mathbf{a} \in \mathbb{R}^c$ and $P \in \mathbb{R}^{n \times h \times w}$ represent the action prediction and the space-time localization map, respectively. The spread loss, denoted as $\mathcal{L}_{\text{cls}}$, is given by

$$\mathcal{L}_{\text{cls}} = \sum_{i \neq \mathbf{y}} \max(0, b - (\mathbf{a}_{\mathbf{y}} - \mathbf{a}_i))^2, \tag{1}$$

where $b \in (0, 1)$ denotes a margin, $\mathbf{a}_i$ is the prediction score for action class i, and $\mathbf{a}_{\mathbf{y}}$ represents that of ground-truth $\mathbf{y}$. The space-time localization loss is a combination of two different losses; the dice loss $\mathcal{L}_{\text{dice}}$ and the cross-entropy loss $\mathcal{L}_{\text{ce}}$. The two losses are defined as follows:

$$\mathcal{L}_{\text{dice}} = \frac{1}{n} \sum_{j=1}^{n} \left(1 - \frac{2 \times \sum_{k=1}^{h \cdot w} (P_{j,k} \cdot \mathcal{Y}_{j,k})}{\sum_{k=1}^{h \cdot w} P_{j,k} + \sum_{k=1}^{h \cdot w} \mathcal{Y}_{j,k}} \right), \tag{2}$$

$$\mathcal{L}_{\text{ce}} = -\frac{1}{n \cdot h \cdot w} \sum_{j=1}^{n} \sum_{k=1}^{h \cdot w} (\mathcal{Y}_{j,k} \log(P_{j,k}) + (1 - \mathcal{Y}_{j,k}) \log(1 - P_{j,k})), \tag{3}$$

where j represents the frame index, k is an index for each pixel of a given frame, and $\mathcal{Y}_{j,k}$ indicates the ground truth of k-th pixel in a j-th frame. The total loss of supervised learning, denoted as $\mathcal{L}_{\text{sup}}$, is the average of aforementioned supervised losses over all the labeled data in D_L, and is given by

$$\mathcal{L}_{\text{sup}} = \frac{1}{|D_L|} \sum_{(X,Y) \in D_L} (\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{ce}}). \tag{4}$$

3.3. Spatio-temporal Semi-supervised Learning

In addition to the supervised learning, our model is also trained with the unlabeled video set D_U by leveraging consistency regularization and contrastive learning in both the spatial and temporal domains. We utilize the mean teacher framework [59], which employs two networks: the student that serves as our main network, and the teacher used to generate pseudo-supervision. The weights $\tilde{\theta}$ of the teacher are updated through an exponential moving average of the weights θ of the student, with an update ratio β:

$$\tilde{\theta}_t = \beta \tilde{\theta}_{t-1} + (1 - \beta)\theta_t. \tag{5}$$

3.3.1 Semi-supervised Learning on Spatial Domain.

The main idea behind semi-supervised learning in the spatial domain is to utilize the spatial consistency within video

frames. Let $E \in \mathbb{R}^{n\times d\times h\times w}$ and $\tilde{E} \in \mathbb{R}^{n\times d\times h\times w}$ denote the latent feature embeddings from the student and teacher networks for an unlabeled video X with an embedding dimension of d, respectively. Also, let cos denote the cosine similarity function. The frame-wise spatial consistency loss, $\mathcal{L}_{\text{scon}}$, is then given by

$$\mathcal{L}_{\text{scon}} = \frac{1}{n\cdot h\cdot w}\sum_{j=1}^{n}\sum_{k=1}^{h\cdot w}\left\{1-\cos(\tilde{E}_{j,k}, E_{j,k})\right\}, \quad (6)$$

where $E_{j,k}, \tilde{E}_{j,k} \in \mathbb{R}^d$ with the frame index j and an index for each pixel of a given frame k. Note that the input fed to the student is augmented differently from that of the teacher. The spatial consistency learning by $\mathcal{L}_{\text{scon}}$ in Eq. (6) improves performance by encouraging the model to make consistent embeddings for the same video in a frame-by-frame manner.

3.3.2 Semi-supervised Learning on Temporal Domain.

In various video understanding tasks including VAD, it is encouraged that the latent feature embeddings for actors and background remain consistent across different clips depicting the same action for robustness to appearance variations and improved discrimination between actors and background in a feature space [60, 65, 75]. However, as will be discussed in Section 4.3, a model trained solely with spatial consistency struggles to maintain robust consistency between such clips and exhibits diminished performance on them as it is trained to achieve consistency of features in the spatial context, neglecting its context variation in the temporal domain.

To address these challenges and promote robust and consistent video representation learning, we introduce a novel temporal data augmentation method, *temporal cross-view augmentation*. This method begins by sampling two clips from the same video, ensuring they share some frames but have different timestamps. Subsequently, consistency regularization is applied to these shared frames. Let X_a and X_b represent video clips sampled from the same video, consisting of n frames, and sharing m frames with each other ($m < n$). The temporal consistency loss, denoted as $\mathcal{L}_{\text{tcon}}$, is given by

$$\mathcal{L}_{\text{tcon}} = \frac{1}{m\cdot h\cdot w}\sum_{j=1}^{m}\sum_{k=1}^{h\cdot w}\left\{1-\cos(\tilde{\mathcal{E}}^a_{j,k}, \mathcal{E}^b_{j,k})\right\}, \quad (7)$$

where $\tilde{\mathcal{E}}^a$ and $\mathcal{E}^b$ are embedding maps of the shared frames from X_a and X_b, respectively. In addition to this, we employ contrastive learning to further utilize unshared frames more effectively. Let $\Phi_p^{i,t}$ denote the set of pixels corresponded to actors with their action class i at t-th frame and $\Phi_n^{i,t}$ denote that of pixels not belonging to the actors. Also, let u represent a similarity function, $u(\mathbf{e}_a, \mathbf{e}_b) = \exp(\cos(\mathbf{e}_a, \mathbf{e}_b)/\tau)$, where τ is a temperature hyper-parameter. The pixel-wise contrastive loss $\mathcal{L}_{\text{cotr}}$ is then given by

$$\mathcal{L}_{\text{cotr}} = -\frac{1}{m\cdot n\cdot c}\sum_{j=1}^{m}\sum_{l=1}^{n}\sum_{i}^{c}\frac{1}{|\Phi_p^{i,j}|}\sum_{k\in\Phi_p^{i,j}} \log\frac{u(\tilde{\mathcal{E}}^a_{j,k}, E^b_{l,k})}{u(\tilde{\mathcal{E}}^a_{j,k}, E^b_{l,k}) + \sum_{o\in\Phi_n^{i,l}} u(\tilde{\mathcal{E}}^a_{j,k}, E^b_{l,o})}. \quad (8)$$

The contrastive loss enhances the performance of the model by learning feature embedding using both shared and unshared frames together. By the temporal consistency learning with $\mathcal{L}_{\text{tcon}}$ in Eq. (7) and $\mathcal{L}_{\text{cotr}}$ in Eq. (8), the model is able to adapt to temporal variations within the same video.

3.3.3 Semi-supervised Action Consistency Learning.

Under the assumption that the model's action predictions on different clips depicting the same action should be the same, the main idea behind action consistency is to utilize spatio-temporal consistency between those clips. The objective of the action consistency learning is to enforce the model to predict the same outcome for (1) clips sampled from the same video with the same timestamps but augmented differently and (2) clips from the same video with different timestamps. With the action consistency learning, the model's ability of recognizing actions is enhanced along with aforementioned spatio-temporal consistency learning. Let X_a and X_b be two clips that hold one of the two conditions above, and $\tilde{\mathbf{a}}^a$ and $\mathbf{a}^b$ be the action predictions for X_a and X_b by the teacher and student network, respectively. The action consistency loss, $\mathcal{L}_{\text{acon}}$, is given by

$$\mathcal{L}_{\text{acon}} = (\tilde{\mathbf{a}}^a - \mathbf{a}^b)^2. \quad (9)$$

3.4. Semi-supervised Learning with Global-Local Context Fusion

Different frames in a video have been known to contain distinct spatio-temporal information that complements each other [30, 37, 49, 61]. Motivated by this, we propose a novel context fusion method, *global-local context fusion*, to enrich features of each frame by utilizing both spatial and temporal context throughout the video. This method extracts latent features from different frames within the same clip, combines them to capture the global context of the clip, and fusing the global context with the local features of each frame.

Let v_s and v_t denote source and target frames, respectively. Also suppose we aim to extract latent features from v_s and fuse them with those of v_t. Initially, we aggregate representative features from the source frame by averaging the features of the actor and those of background regions separately. Let $E_j \in \mathbb{R}^{d\times h\times w}$ be a feature map and

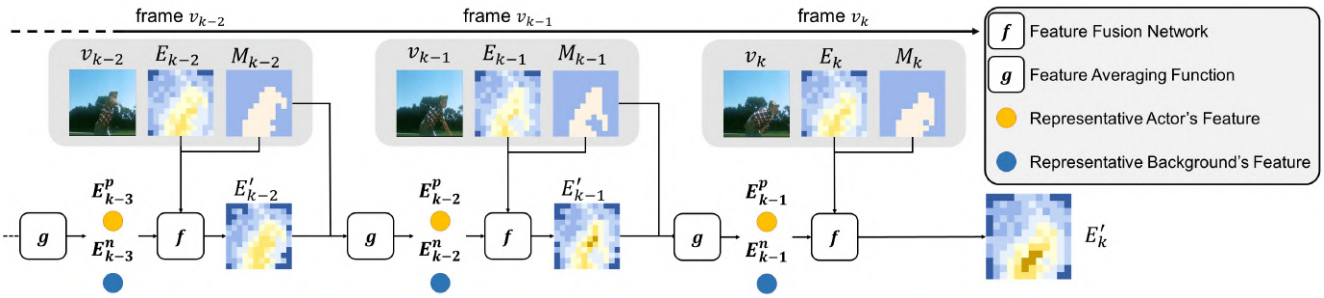

Figure 2. Process of the global-local context fusion. Through the feature fusion network f, each feature embedding is fused with the preceding representative actor and background features, which are generated by the feature averaging function g.

$M_j \in \{0,1\}^{h\times w}$ be a pseudo localization map obtained from the outputs of the unlabeled video, where 1 indicates the regions the action occurs. Let us define an averaging function g by

$$g(E_j, M_j) = \frac{\sum_{k=1}^{h\cdot w} M_{j,k} \cdot E_{j,k}}{\sum_{k=1}^{h\cdot w} M_{j,k}}. \tag{10}$$

The representative actor and background features of v_s are then computed as follows: $E^p_s = g(E_s, M_s)$ and $E^n_s = g(E_s, \neg M_s)$, where E_s denotes the latent features of v_s and $M_s \in \{0,1\}^{h\times w}$ denotes a pseudo localization map from the source frame. Next, we propagate the aggregated features, E^p_s and E^n_s, from the source frame v_s to the target frame v_t. Let E_t and M_t be a feature embedding and pseudo localization map of the target frame v_t, respectively. Then the fused feature of the target frame is given by

$$E'_{t,k} = \begin{cases} f(E^p_s \oplus E_{t,k}) & \text{if } M_{t,k} = 1, \\ f(E^n_s \oplus E_{t,k}) & \text{otherwise,} \end{cases} \tag{11}$$

where $\oplus$ represents channel-wise matrix concatenation and f is a simple feature fusion network that takes input in $\mathbb{R}^{2d}$ and returns the output in $\mathbb{R}^d$.

Now suppose we are given a sequence of k frames $(v_1, v_2, ..., v_k)$ and our goal is to enhance features of all the frames via the iterative global-local context fusion process from v_1 to v_k. Specifically, the fusion process begins by operating on the first two frames, v_1 and v_2, which represent source and target frames, respectively, and continues by iteratively applying the context fusion on the next consecutive frame pair (v_2 and v_3, and so on) till the last frame pair (v_{k-1}, v_k). Note that at each iteration, the target frame v_t and its fused feature E'_t from the previous iteration now become the source frame v_s and its feature embedding E_s for the next iteration. This process is depicted in Fig. 2.

Yet regarding this fusion process, we should carefully address the following consideration: How do we select a suitable sequence of frames for the process? It is infeasible to identify the optimal sequence out of all possible candidates due to a prohibitively large number of such sequences and expensive post-hoc evaluation. Hence, we consider all possible sequences complying certain conditions and refer to each of these sequences as a *path*, which then represents a sequential fusion process. This strategy allows to handle uncertainty inherent in the sequence selection for the fusion process, like sequential model averaging techniques [16, 55]. Let r be the number of frames within a sequence that comprises a single path, and j an index for the last target frame of the fusion process. Then, the set of all possible paths Ω_j to the frame v_j is defined by

$$\Omega_j = \bigcup_{r=2}^{n} \Big\{ p \,|\, p = [v_{k_1}, ..., v_{k_r}], \\ (1 \le k_1 < ... < k_r = j) \vee (n \ge k_1 > ... > k_r = j) \Big\}, \tag{12}$$

where p denotes a single path in the form of a list, and it consists of frames sorted in a monotonic increasing or decreasing order on the time axis. Then we randomly sample s paths from Ω_j per last target frame to achieve temporal drop-out regularization with computation efficiency. Note that the time complexity of the context fusion is $O(ns|p|)$, and $|p|$ is empirically lower than n (*i.e.*, $|p| < n$).

Finally, for the last target frame v_j, the global-local context fusion is performed for each path of Ω_j to obtain a fused feature embedding of the target frame, and the final fused feature embedding is obtained by *averaging all the fused feature embeddings* from every path. Note that we apply the context fusion to the feature embeddings of the teacher network and then perform the proposed spatio-temporal consistency and contrastive learning with those of the teacher network and raw feature embeddings from the student network.

The overall semi-supervised learning loss, $\mathcal{L}_{\text{semi}}$, is computed as the average of the aforementioned losses over all

the unlabeled data in D_U and is given by:

$$\mathcal{L}_{\text{semi}} = \frac{1}{|D_U|} \sum_{X \in D_U} (\mathcal{L}_{\text{scon}} + \mathcal{L}_{\text{tcon}} + \mathcal{L}_{\text{cotr}} + \mathcal{L}_{\text{acon}}). \quad (13)$$

The final objective is a combination of $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{semi}}$, and is given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda \cdot \mathcal{L}_{\text{semi}}, \quad (14)$$

where $\lambda \in (0, 1]$ is a Gaussian ramp-up constant [31], which gradually increases as the training progresses.

4. Experiments

4.1. Implementation Details

Datasets. We conducted experiments on two datasets, UCF101-24 [57] and JHMDB-21 [21]. UCF101-24 comprises videos of 24 action classes, where about 78% of the videos are untrimmed. It presents 2,284 videos for training and 923 for testing. Meanwhile, JHMDB-21 consists of all trimmed videos of 21 action classes; 660 videos are used for training and the other 268 videos are kept for testing. For training on both UCF101-24 and JHMDB-21, we use a frame resolution of 224 × 224. The number of frames per clip, n, is set to 8 with a temporal skip rate of 2. For a fair comparison with previous work, we split the training set of UCF101-24 and JHMDB-21 into labeled and unlabeled sets with a ratio of 30% and 20%, respectively, following [25].

Network architecture. In line with [25,72] for UCF101-24 and JHMDB-21, we utilize VideoCapsuleNet [9] with I3D backbone [5] as our VAD network. We replace 3D routing with 2D routing [45], leading to improved efficiency in terms of memory usage and training speed. The backbone is pre-trained on Kinetics [23] and Charades [48] dataset. We employ an additional module for feature embeddings, implemented with a single 3D convolutional layer. The feature fusion network f used in Eq. (11) consists of Conv2D-BN-ReLU-Conv2D modules in sequence.

Data augmentation. We employ data augmentation methods to enhance traning within our framework, such as Gaussian blurring, color-jittering, grayscaling, and horizontal flipping.

Optimizer. We use AdamW [35] with learning rate 5e-5 and weight decay 1e-6.

Hyper-parameters. For labeled data, the mini-batch size is 8, while it is 2 for unlabeled data on both UCF101-24 and JHMDB-21 datasets. The margin value in Eq.(1) is set to 0.2, following [25]. The update ratio β is set to 0.995. The temperature value τ for $\mathcal{L}_{\text{cotr}}$ is set to 0.1. The number of shared frames in Section 3.3.2, m, is set to 4. The number of sampled paths for each target frame denoted as s, is set to 3. The Gaussian ramp-up constant in Eq. (14) is calculated as $\min(\exp(-0.05 \times (100 - t)), 1)$, where t is a current epoch. More details about hyper-parameters are found in the Supplement Materials.

Method	Net.	UCF101-24 f-mAP 0.5	UCF101-24 v-mAP 0.2	UCF101-24 v-mAP 0.5	JHMDB-21 f-mAP 0.5	JHMDB-21 v-mAP 0.2	JHMDB-21 v-mAP 0.5
Pseudo-label† [32]	VDCap.	64.9	93.0	65.6	57.4	90.1	57.4
MixMatch† [3]	VDCap.	20.2	60.2	13.8	7.5	46.2	5.8
Co-SSD(CC)† [20]	VDCap.	65.3	93.7	67.5	60.7	94.3	58.5
Kumar *et al.*‡ [25]	VDCap.	71.5	96.5	73.5	65.9	97.2	68.0
BWCC† [72]	VDCap.	71.9	95.3	73.7	66.3	94.8	68.1
Ours	VDCap.	**80.0**	**97.1**	**82.5**	**81.8**	**98.5**	**83.3**
Full	VDCap.	82.6	98.1	83.3	85.8	99.5	84.9

Table 1. Comparisons with semi-supervised methods using various thresholds for f-mAP and v-mAP on UCF101-24 and JHMDB-21 test sets. Note that action prediction was not considered in this evaluation. † denotes performances obtained from [25, 72]. ‡ denotes a re-implemented method.

Method	Net.	UCF101-24 f-mAP 0.5	UCF101-24 v-mAP 0.2	UCF101-24 v-mAP 0.5	JHMDB-21 f-mAP 0.5	JHMDB-21 v-mAP 0.2	JHMDB-21 v-mAP 0.5
Kumar *et al.*‡ [25]	VDCap.	58.3	80.7	61.3	27.7	37.9	29.6
Ours	VDCap.	**68.8**	**83.5**	**71.5**	**34.9**	**38.2**	**35.3**
Full	VDCap.	72.0	86.1	73.3	47.1	48.1	42.7

Table 2. Comparisons with the semi-supervised method of [25] using various thresholds for f-mAP and v-mAP on UCF101-24 and JHMDB-21 test sets. Note that action prediction was considered in this evaluation. ‡ denotes a re-implemented method.

Evaluation metric. We calculate spatial IoU for each frame to determine frame average precision and compute spatio-temporal IoU per video for video average precision. We then average these scores to obtain f-mAP and v-mAP over various thresholds. In our evaluation, a correct prediction requires matching both the predicted action label and action localization maps with the ground truth. Some previous work [25, 72] did not include the predicted action label in their evaluation. We will report both metrics with and without the action label.

4.2. Results

Comparison with the state-of-the-art without the action label. We compare our method to state-of-the-art semi-supervised VAD [3, 20, 25, 32, 72] and the result of training in a fully supervised manner (notated as Full). We conducted experiments on UCF101-24 and JHMDB-21. We abbreviate VideoCapsuleNet [9], the base model for all listed semi-supervised methods, as VDCap, for brevity. Results without the action label are reported in Table 1, demonstrating that our method achieves state-of-the-art performance for all evaluation settings. For UCF101-24, our method shows significant performance improvement over previous methods, achieving 97.1 and 82.5 in v-mAP@0.2 and v-mAP@0.5, respectively. Similarly, for JHMDB-21, our method surpasses all previous work by a large margin, showing 98.5 and 83.3 in the threshold of v-mAP@0.2 and

ID	$\mathcal{L}_{\text{sup}}$	$\mathcal{L}_{\text{scon}}$	$\mathcal{L}_{\text{tcon}}$	$\mathcal{L}_{\text{cotr}}$	$\mathcal{L}_{\text{acon}}$	GLF	UCF101-24				JHMDB-21			
							f-mAP		v-mAP		f-mAP		v-mAP	
							0.2	0.5	0.2	0.5	0.2	0.5	0.2	0.5
I	✓						90.7 / 75.5	67.4 / 57.6	95.3 / 79.4	68.5 / 58.2	87.5 / 35.1	63.5 / 25.9	94.8 / 31.9	64.2 / 28.0
II	✓		✓		▲		90.9 / 77.7	77.6 / 66.4	96.5 / 79.9	79.6 / 68.6	95.6 / 36.2	74.1 / 31.7	98.0 / 36.6	77.3 / 32.3
III	✓		✓	✓	▲		91.7 / 78.3	79.5 / 68.0	96.8 / 82.9	80.9 / 70.3	96.0 / 37.5	79.8 / 33.1	98.2 / 37.7	80.5 / 32.7
IV	✓		✓		▲	✓	91.1 / 78.4	77.9 / 67.2	96.7 / 81.3	80.2 / 69.2	95.9 / 36.5	78.8 / 32.6	98.1 / 36.8	80.2 / 32.9
V	✓		✓	✓	▲	✓	92.4 / 79.6	79.5 / 68.5	97.0 / 83.3	82.1 / 70.8	96.8 / 37.9	81.7 / 34.0	98.3 / 38.1	82.0 / 34.8
VI	✓	✓			▲		90.8 / 77.3	77.2 / 66.6	96.8 / 80.5	78.4 / 66.5	95.3 / 35.7	73.2 / 31.7	97.8 / 36.8	76.0 / 31.9
VII	✓	✓	✓		✓		91.0 / 78.2	77.6 / 66.7	96.8 / 81.5	80.0 / 68.8	95.8 / 36.4	75.8 / 32.8	98.4 / 37.0	78.9 / 32.7
VIII	✓	✓	✓	✓	✓		92.3 / 79.1	79.7 / 68.3	96.9 / 83.0	81.5 / 71.2	96.1 / 38.0	80.8 / 34.0	98.4 / 37.8	82.0 / 32.8
IX	✓	✓			▲	✓	91.5 / 77.9	77.9 / 68.7	97.0 / 82.0	81.2 / 70.9	96.1 / 36.2	77.7 / 32.8	98.0 / 37.3	80.5 / 33.2
X	✓	✓	✓	✓	✓	✓	**92.5 / 79.8**	**80.0 / 68.8**	**97.1 / 83.5**	**82.5 / 71.5**	**97.0 / 38.2**	**81.8 / 34.9**	**98.5 / 38.2**	**83.3 / 35.3**

Table 3. Ablation studies of losses and global-local context fusion (GLF) in Eq. (13) on UCF101-24 and JHMDB-21 test sets. In $\mathcal{L}_{\text{acon}}$, ▲ indicates that the loss function considers the consistency between clips with either the same or different timestamps, but not both. Performance is reported without the action label, then with the label included (separated by '/').

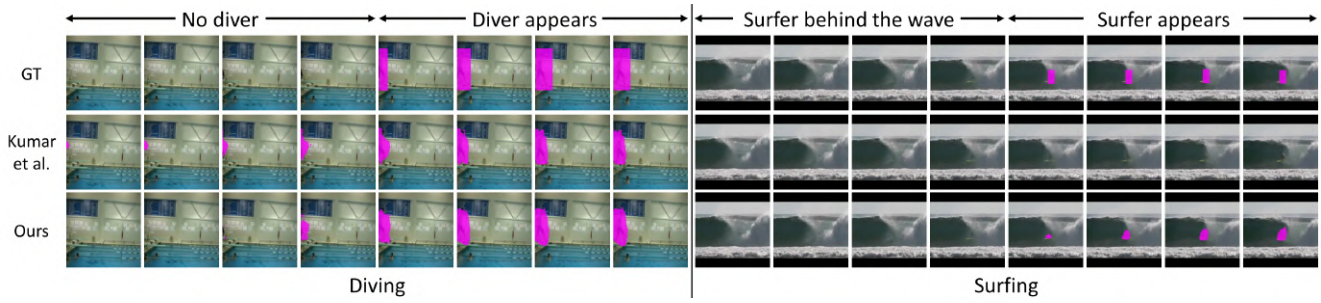

Figure 3. Qualitative results on *test* set of UCF101-24 [57] with ours and Kumar *et al.* [25]

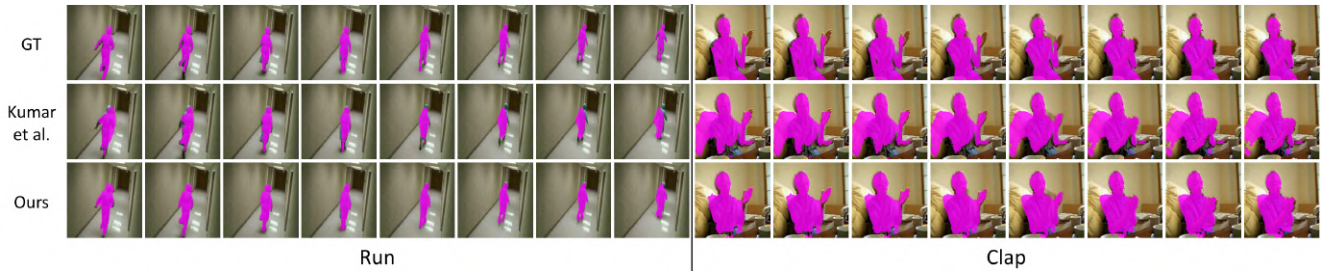

Figure 4. Qualitative results on *test* set of JHMDB-21 [21] with ours and Kumar *et al.* [25]

v-mAP@0.5, respectively.

Comparison with the state-of-the-art with the action label. The results with the action label are shown in Table 2. Previous work [25, 72] and other reported methods did not include the action label in their evaluation protocol. To address this, we re-implemented [25] as it was the only one with its official codebase and reported the performance. The results in Table 2 show that our method consistently outperforms the previous methods.

Qualitative analysis. Fig. 3 displays our results on UCF101-24, while Fig. 4 exhibits those on JHMDB-21.

4.3. Ablation Studies

We conducted ablation studies to investigate the effectiveness of each component of the proposed method. The experiments were conducted on test sets of UCF101-24 [57] and JHMDB-21 [21].

Impact on losses and global-local context fusion in Eq. (13). We investigated the impact of each loss term and global-local context fusion. The results are showcased in Table 3. The experiments I, II and III show that the effectiveness of the temporal-consistency learning (Exp. II) compared to the supervised only (Exp. I), and contrastive learning (Exp. III) improves its performance based on temporal-consistency learning (Exp. II). For the global-local context fusion, the experiments II, III, VI and IV, V, IX show that feature embeddings enhanced by the context fusion (Exp. IV, V and IX) show the improved performance than those not enhanced (Exp. II, III and VI). Note that our method achieved comparable performance with the compared methods by consistency based losses ($\mathcal{L}_{\text{scon}}$ and $\mathcal{L}_{\text{tcon}}$) thanks to our augmentation, training scheduling and optimization strategies. Finally, the best results are achieved when all the components are employed.

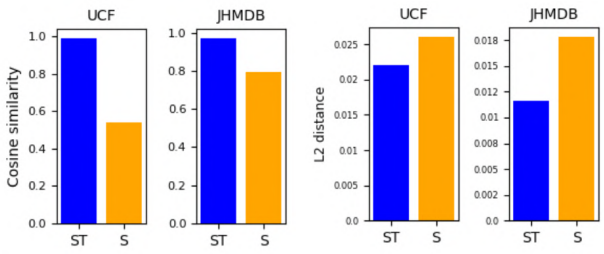

Figure 5. Ablation studies of consistency on the shared frames from two clips. (*left*) Averaged cosine similarity of feature embeddings on shared frames. (*right*) Averaged L2-distance of action localization maps on shared frames.

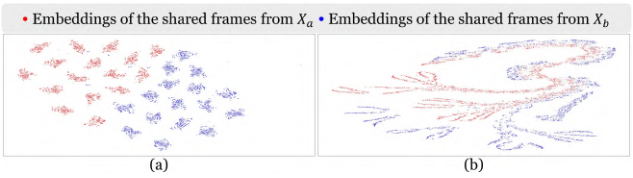

Figure 6. t-SNE visualization of embedding distributions of frames shared by overlapping clips X_a and X_b. (a) Result of a model trained without temporal consistency. (b) Result of a model trained with temporal consistency.

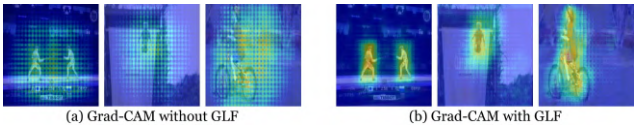

Figure 7. Grad-CAM [47] visualization of spatial consistency loss without and with GLF.

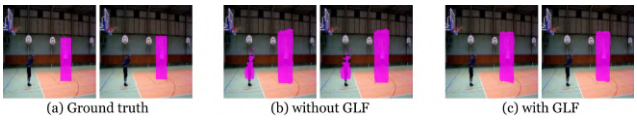

Figure 8. Qualitative results on the UCF101-24 *test* set without and with GLF.

Robust consistency on shared frames. We noticed that a model trained only on spatial consistency struggles to maintain consistency on shared frames from clips of the same video. To investigate, we extracted two clips with shared frames from the same video and conducted an experiment; we compared a model trained with temporal consistency losses ($\mathcal{L}_{\text{tcon}}$ and $\mathcal{L}_{\text{cotr}}$) in addition to spatial consistency loss ($\mathcal{L}_{\text{scon}}$) (denoted as **ST**) against a model trained solely with spatial consistency loss ($\mathcal{L}_{\text{scon}}$) (denoted as **S**). For shared frames, we computed cosine similarity for features and L2-distance for localization maps. High similarity indicates consistent features, while low L2-distance suggests consistent localization predictions. Results (see Fig. 5) demonstrate that the model trained with proposed temporal consistency losses exhibit robust consistency.

Interpretation of the impact of temporal information.

Metric	TubeR	TubeR + Kumar *et al.*	TubeR + Ours	Full
f-mAP@0.5	16.5	18.3	**22.5**	28.8
Metric	STMixer	STMixer+ Kumar *et al.*	STMixer + Ours	Full
f-mAP@0.5	22.4	23.9	**25.7**	27.2

Table 4. Additional experiments on AVA, where we compare our model and Kumar *et al.* [25] incorporated with TubeR [71] and STMixer [63], respectively.

We studied the impact of temporal information in temporal consistency and GLF. For temporal consistency, we first trained two models, one with and the other without temporal consistency; we then visualized their embedding distributions for frames shared by two overlapping clips. As in Fig. 6, the model with temporal consistency showed aligned embedding distributions despite different temporal contexts of the clips; this leads to learned representation robust to temporal variations. We also visualized the impact of GLF during training through Grad-CAMs [47] of the spatial consistency loss with and without GLF in Fig. 7. The model with GLF focused its attention more on the actors of interest, reducing false action predictions and enhancing performance, as also depicted in Fig. 8.

4.4. Additional experiments on AVA

AVA [14] is a large-scale dataset comprising 299 movies, each lasting 15 minutes. To showcase the versatility and superiority of our method, we conducted additional experiments on AVA. Specifically, we compared our model with Kumar *et al.* [25], integrating TubeR [71] and STMixer [63] as a base VAD model in the semi-supervised setting, where only 10% of the data is labeled. Following the training and evaluation protocol outlined in TubeR and STMixer, results in Table 4 demonstrate the effectiveness and generalization ability of our method.

5. Conclusion

We have presented a new semi-supervised learning framework for video action detection, incorporating a novel temporal augmentation strategy and global-local context fusion. To mitigate the discrepancy in feature embeddings between two clips from the same video and more effectively utilize temporal context during training, we introduce a novel temporal-augmentation method called *temporal cross-view augmentation*, along with global-local context fusion. We then train a model with consistency regularization and contrastive learning using both proposed components. Our framework substantially improves performance over existing semi-supervised methods for all the benchmarks.

Acknowledgment. This work was supported by Samsung Electronics Co., Ltd (IO201210-07948-01) and the NRF grant funded by Ministry of Science and ICT, Korea (NRF-2021R1A2C3012728).

References

[1] Iñigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 3

[2] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 3

[3] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Proc. Neural Information Processing Systems (NeurIPS)*, 32, 2019. 6

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[5] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 6

[6] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13638–13647, 2021. 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[8] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 1060–1068, January 2021. 1

[9] Kevin Duarte, Yogesh Rawat, and Mubarak Shah. Videocapsulenet: A simplified network for action detection. *Advances in neural information processing systems*, 31, 2018. 1, 2, 3, 6

[10] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *Proc. British Machine Vision Conference (BMVC)*, 2020. 1

[11] Harshala Gammulle, David Ahmedt-Aristizabal, Simon Denman, Lachlan Tychsen-Smith, Lars Petersson, and Clinton Fookes. Continuous human action recognition for human-machine interaction: A review. *ACM Comput. Surv.*, 55(13s), jul 2023. 1

[12] Yongbin Gao, Xuehao Xiang, Naixue Xiong, Bo Huang, Hyo Jong Lee, Rad Alrifai, Xiaoyan Jiang, and Zhijun Fang. Human action monitoring for healthcare based on deep learning. *IEEE Access*, 6:52277–52285, 2018. 1

[13] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 759–768, 2015. 2

[14] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 1, 2, 8

[15] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. 3

[16] Seunghoon Hong, Suha Kwak, and Bohyung Han. Orderless tracking through model-averaged posterior estimation. In *2013 IEEE International Conference on Computer Vision*, pages 2296–2303, 2013. 5

[17] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 5822–5831, 2017. 1, 2

[18] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 1

[19] Ahmad Jalal, Shaharyar Kamal, and Daijin Kim. A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. *Sensors*, 14(7):11735–11759, 2014. 1

[20] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32, 2019. 3, 6

[21] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *2013 IEEE International Conference on Computer Vision*, pages 3192–3199, 2013. 1, 2, 6, 7

[22] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4405–4413, 2017. 1, 2

[23] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017. 6

[24] Rihuan Ke, Angelica Aviles-Rivero, Saurabh Pandey, Saikumar Reddy, and Carola-Bibiane Schönlieb. A three-stage self-training framework for semi-supervised semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[25] Akash Kumar and Yogesh Singh Rawat. End-to-end semi-supervised learning for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14700–14710, 2022. 1, 3, 6, 7, 8

[26] Vijay Kumar B G, Gustavo Carneiro, and Ian Reid. Learning local image descriptors with deep siamese and triplet

convolutional networks by minimising global loss functions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[27] Donghyeon Kwon, Minsu Cho, and Suha Kwak. Self-supervised learning of semantic correspondence using web videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2142–2152, January 2024. 3

[28] Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9957–9967, June 2022. 1, 3

[29] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[30] Yuandu Lai, Yahong Han, and Yaowei Wang. Anomaly detection with prototype-guided discriminative latent embeddings. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 300–309, 2021. 2, 4

[31] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 6

[32] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013. 6

[33] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 68–84. Springer, 2020. 1, 2

[34] Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge. In *ICCV*, 2023. 2

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. International Conference on Learning Representations (ICLR)*, 2019. 6

[36] Rajat Modi, Aayush Jung Rana, Akash Kumar, Praveen Tirupattur, Shruti Vyas, Yogesh Rawat, and Mubarak Shah. Video action detection: Analysing limitations and challenges. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4911–4920, June 2022. 1

[37] Rajat Modi, Aayush Jung Rana, Akash Kumar, Praveen Tirupattur, Shruti Vyas, Yogesh Rawat, and Mubarak Shah. Video action detection: Analysing limitations and challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4911–4920, 2022. 2, 4

[38] Inzamam Mashood Nasir, Mudassar Raza, Jamal Hussain Shah, Muhammad Attique Khan, and Amjad Rehman. Human action recognition using machine learning in uncontrolled environment. In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, pages 182–187, 2021. 1

[39] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xioyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, and Mita Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3153–3160, 2011. 1

[40] Yassine Ouali, Celine Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[41] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 464–474, 2021. 1

[42] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream r-cnn for action detection. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 744–759. Springer, 2016. 1

[43] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G Hauptmann. Rethinking zero-shot action recognition: Learning from latent atomic actions. In *European Conference on Computer Vision*, pages 104–120. Springer, 2022. 2

[44] Ashwin Ramachandran, Kartik Gokhale, Maike Kripps, and Thomas Deserno. Video-based in-vehicle action recognition for continuous health monitoring. In *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*, volume 12469, pages 197–210. SPIE, 2023. 1

[45] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3859–3869, Red Hook, NY, USA, 2017. Curran Associates Inc. 6

[46] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. *arXiv preprint arXiv:1608.01529*, 2016. 1, 2

[47] 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. 8

[48] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 510–526, Cham, 2016. Springer International Publishing. 6

[49] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Pro-*

ceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1, NIPS'14, page 568–576, Cambridge, MA, USA, 2014. MIT Press. 2, 4

[50] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10399, 2021. 3

[51] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10389–10399, June 2021. 3

[52] Ayush Singh, Aayush J Rana, Akash Kumar, Shruti Vyas, and Yogesh Singh Rawat. Semi-supervised active learning for video action detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4891–4899, 2024. 3

[53] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 3637–3646, 2017. 1, 2

[54] 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 *Proc. Neural Information Processing Systems (NeurIPS)*, 2020. 1, 3

[55] Jeany Son. Contrastive learning for space-time correspondence via self-cycle consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14679–14688, June 2022. 5

[56] Lin Song, Shiwei Zhang, Gang Yu, and Hongbin Sun. Tacnet: Transition-aware context network for spatio-temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11987–11995, 2019. 2

[57] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1, 2, 6, 7

[58] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–334, 2018. 1

[59] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2017. 3

[60] Anyang Tong, Chao Tang, and Wenjian Wang. Semi-supervised action recognition from temporal augmentation using curriculum learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1305–1319, 2023. 4

[61] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2740–2755, 2019. 2, 4

[62] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision*, pages 3164–3172, 2015. 2

[63] Tao Wu, Mengqi Cao, Ziteng Gao, Gangshan Wu, and Limin Wang. Stmixer: A one-stage sparse action detector. In *CVPR*, 2023. 8

[64] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 3

[65] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. Learning from temporal gradient for semi-supervised action recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 4

[66] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2015. 2

[67] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 3

[68] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 264–272, 2019. 1, 2

[69] Hongcheng Zhang, Xu Zhao, and Dongqi Wang. Semi-supervised learning for multi-label video action detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 2124–2134, New York, NY, USA, 2022. Association for Computing Machinery. 3

[70] Jiaojiao Zhao and Cees GM Snoek. Dance with flow: Two-in-one stream action detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 9935–9944, 2019. 1, 2

[71] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. Tuber: Tubelet transformer for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13598–13607, 2022. 8

[72] Xian Zhong, Aoyu Yi, Wenxuan Liu, Wenxin Huang, Chengming Zou, and Zheng Wang. Background-weakening consistency regularization for semi-supervised video action detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1, 3, 6, 7

[73] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan2, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent

semi-supervised semantic segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[74] Yanning Zhou, Hang Xu, and Wei Zhang. c^3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[75] Yuliang Zou, Jinwoo Choi, Qitong Wang, and Jia-Bin Huang. Learning representational invariances for data-efficient action recognition. *Comput. Vis. Image Underst.*, 227(C), jan 2023. 4