Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2023) 31: 1129 – 1145

doi:10.55730/1300-0632.4038

Feature distillation from vision-language model for semisupervised action classification

Aslı ÇELİK, Ayhan KÜÇÜKMANİSA*, Oğuzhan URHAN
Department of Electronics and Communications Engineering, Faculty of Engineering, Kocaeli University, Kocaeli, Turkiye

Received: 07.07.2023 • **Accepted/Published Online:** 01.10.2023 • **Final Version:** 27.10.2023

Abstract: The training of supervised machine learning approaches is critically dependent on annotating large-scale datasets. Semisupervised learning approaches aim to achieve compatible performance with supervised methods using relatively less annotation without sacrificing good generalization capacity. In line with this objective, ways of leveraging unlabeled data have been the subject of intense research. However, semisupervised video action recognition has received relatively less attention compared to image domain implementations. Existing semisupervised video action recognition methods trained from scratch rely heavily on augmentation techniques, complex architectures, and/or the use of other modalities while distillation-based methods use models that have only been trained for 2D computer vision tasks. In another line of work, pretrained vision-language models have shown very promising results for generating general-purpose visual features with reports of high zero-shot performance for many downstream tasks. In this work, we exploit a language-supervised visual encoder for learning video representations for video action classification tasks. We propose a teacher-student learning paradigm through feature distillation and pseudo-labeling. Our experimental results are a proof-of-concept revealing that multimodal feature extractors can be utilized for spatiotemporal feature extraction in a semisupervised learning context and show compatible performance with SOTA methods, especially in a low-label regime.

Key words: Video action classification, multimodal learning, semisupervised learning, feature distillation

1. Introduction

Supervised deep learning approaches for video action classification have yielded great success [1–6]. Unfortunately, the training of supervised approaches is critically dependent on annotating large-scale datasets. For tasks of video understanding, there is a high cost of data annotation. Despite the availability of video data on the internet, curating videos and creating annotations is still expensive. Moreover, training with large-scale datasets is time- and resource-consuming. Semisupervised learning (SSL) approaches aim to achieve compatible performance with supervised methods using relatively less annotation without sacrificing good generalization capacity [7–10]. In this direction, making use of data without labeling is an intense topic of recent research. However, the use of SSL approaches in video action recognition has received relatively less attention.

Natural language supervision has been utilized for visual representation learning and proven to be very effective for generating joint semantic embedding space [11, 12]. In terms of data acquisition convenience, using raw text data as a training signal is advantageous compared to classical machine learning annotation formats. For instance, the CLIP [12] backbone is trained on a dataset of (image, text) pairs collected from a variety of

*Correspondence: ayhan.kucukmanisa@kocaeli.edu.tr

publicly available sources on the internet. Generated embedding space has also been proven to be very effective for zero-shot inference in different downstream vision tasks [13–15].

Inspired by promising work on zero-shot transfer, natural language supervision, and the SSL literature, we propose to transfer/distill knowledge from semantic multimodal joint embedding space for better SSL implementations. In this work, we propose to exploit a frozen text-pretrained visual encoder backbone for video representation learning. Given the reported zero-shot performances, our goal is to leverage strong action-related information content in multimodal representations, which will lead to training better models for predicting action classes given a video. Architecture details and the dataset used in the training of the mentioned transformer-based space-time encoder can be found in [16]. We perform feature distillation pretraining using mean square error (MSE) loss by leveraging video data without using any labels. The feature distillation pretraining stage is followed by a fine-tuning stage of both the teacher and student network by introducing cross-entropy loss calculated for the labeled portion of the training data and pseudo-cross-entropy loss calculated for training data pseudo-labeled by the teacher network.

The main contributions of this paper can be summarized as follows:

- We used large-scale multimodal training as an auxiliary for feature distillation in contrast to previous methods that use pretrained fixed weight still-image networks.
- Vision-language models are proven to be strong fine-tuners. In this work, learned knowledge is transferred to a customized network leveraging a simple teacher-student training scheme.

The organization of this manuscript is as follows. Section 2 introduces related studies on semisupervised action recognition, vision-language models, and knowledge distillation. Section 3 presents the method proposed in this work. Section 4 provides experimental results. The evaluation of the findings, limitations of the study, and future research directions are discussed in Section 5. Finally, we summarize our findings in Section 6.

2. Related work

It is possible to examine related studies within the scope of this work within the three categories of semisupervised action recognition, vision-language models, and knowledge distillation. The subsection on semisupervised action recognition first describes deep learning-based architectures and algorithms in the literature that are primarily used for action recognition tasks. The SSL frameworks and existing semisupervised action recognition methods in the literature are then explained. In the subsection on vision-language models, natural language supervision, multimodal joint embedding space, and zero-shot action classification are explained. The subsection on knowledge distillation explains the concept of knowledge distillation in the context of deep learning.

2.1. Semisupervised action recognition

Video action understanding has been extensively studied in terms of action recognition [1, 4, 6, 44–46]. Action recognition is the task of identifying and classifying human actions or activities from video data, with potential applications in various fields. Action recognition involves feature extraction and video-level prediction. In computer vision, deep convolutional neural network (CNN)-based architectures were predominantly used for visual recognition tasks. Conventional methods utilize two-stream CNNs to process temporal and spatial information separately with RGB frames and optical flow [17, 35]. After the introduction of 3D CNNs with 3D convolutions, they can learn spatial and temporal information together in videos for better representation

learning since spatiotemporal video volumes can be processed in CNN-based architectures such as C3D [6], I3D [1], SlowFast [4], X3D [3], and ResNet3D [19]. In order to avoid computational costs, 2D CNNs have also been utilized with additional temporal modules [47, 48, 50, 51]. Due to inspiration from the success of transformer-based architectures in natural language processing, there is also a shift from CNN-based architectures to transformer-based ones in the computer vision community with the introduction of vision transformers [18]. Parallel to this work, action recognition models based on vision transformers have also been proposed recently [38–40, 52]. However, annotation labor is the main disadvantage of these methods, which depend only on supervised training. Pretraining on large datasets followed by fine-tuning of the target dataset also became a common practice after the introduction of large-scale datasets like Kinetics or Sports-1M [36].

The main objective of SSL is constructing models that utilize unlabeled data in conjunction with labeled data to improve performance, especially in cases where large volumes of unlabeled data are available but labeling is challenging, expensive, and/or not feasible. Pseudo-Label [31] and Mean-Teacher [9] are baseline SSL frameworks proposed for the image domain originally. Pseudo-Label assigns maximum predicted labels as if they were true labels for unlabeled samples and trains the network in a supervised manner. Mean-Teacher trains two identical networks called 'teacher' and 'student' simultaneously using cross-entropy loss and consistency regularization loss. Cross-entropy loss is calculated for student network predictions for labeled samples, whereas consistency regularization loss is calculated between student and teacher predictions for noise-applied samples for minimizing the difference in predictions. The weights of the teacher network are updated as an exponential moving average of the weights of the students. S4L [33] unifies SSL with self-supervised learning. Self-supervised training initially trains a network for a pretext task and learned representations improve the performance in downstream tasks as well [53]. Specifically, in [33], a network was trained for the pretext task of predicting applied rotation to an image.

Although SSL in the 2D image domain has been proven to be very effective, there is relatively less work on semisupervised action recognition. Recent studies proposed the usage of regulatory signals from fixed/frozen pretrained networks for knowledge distillation through feature consistency. For example, VideoSSL [20] allowed the exploitation of predictions of a 2D image classifier CNN to distill the information related to the objects of interest in the video based on the assumption that the appearance of objects can be an indication of the actions that take place in the video clip. To do this, soft cross-entropy loss is used, which treats the predictions of a 2D ResNet trained for an image classification task as soft labels. Similarly, DANet [21] leverages multiple auxiliary networks pretrained for static-image computer vision tasks. In this work, positive and negative video pairs are created (positive meaning from the same video, versus negative meaning from different videos) and a weighted contrastive loss is used for feature consistency. In another line of work, [22] suggested a two-pathway temporal contrastive model and processed unlabeled videos at two different speeds leveraging the consistency assumption that changing the speed of the video does not change the action. ActorCutMix [34] was used for a video data augmentation strategy for scene invariance since action recognition datasets show scene biases causing models to focus more on the scene rather than the action itself [54]. Proposed augmentations are then plugged into SSL frameworks such as UDA [55] and FixMatch [7] for data-efficient action recognition. Learning2Augment [56] was used to propose a video augmentation method for composite videos. Based on the predictions of the selector network, video pairs to be used for augmentation are picked and novel videos are created through video compositing. Specifically, one of the pairs is used as the foreground and the other as the background. After background-foreground segmentation, novel videos are created through image inpainting. TACL [49] was also

used to propose temporal action augmentation for extracting coarse and fine-grained action representations from videos and a semisupervised action consistency learning framework for dynamic threshold evaluation in pseudo-labeling. In [23], cross-model pseudo-labeling (CMPL) was suggested, where an auxiliary backbone with different depths is used to get a complementary representation for better pseudo-labeling. In [24], it was proposed to utilize a temporal gradient for additional modality for better temporal information encoding, and [32] also used a temporal gradient in addition to optical flow information. Recently, SVFormer [41] was proposed for the use of transformer models and introduction of augmentations for spatial and temporal domains.

2.2. Vision-language models

Natural language can be both a supervision and a prediction space. Raw text data can be utilized as an alternative training signal to classical machine learning annotations. Images can be mapped into the semantic space of words and this flexible prediction space enables zero-shot learning, which is the ability to classify instances of a class that is not seen. The method presented in [25] is a proof-of-concept for the mentioned cross-modal transfer. For learning visual representations, CLIP [12] trains an image and text encoder in order to learn correct pairings for training data using contrastive learning. Since the introduction of the CLIP backbone, vision-language models have been candidates for vision foundation models. 'Foundation model' here refers to a task-agnostic model that generates general-purpose visual features for any downstream task. There is a growing body of literature on multimodal learning combining vision and language modalities for generating joint embedding spaces for the image domain ALIGN [57] and Florence [58], as well as the video domain Violet [59], ClipBert [37], and Frozen in Time [16]. Parallel to this work, large-scale video datasets have also been introduced [16, 42, 43].

Zero-shot classification inference in generated embedding space reduces to searching and finding the test class whose embedding is the nearest-neighbor of the model's feature extractor output. Figure 1 illustrates a representation of the vision-language joint embedding space. During training of a joint embedding space, images/videos and associated textual descriptions or captions are presented to the model. The visual and textual features are extracted with a visual and text encoder, respectively. Minimizing the distance between pairs of visual and textual representations while maximizing the distance to other nonmatching pairs is the training objective. In this way, the model learns to recognize/capture the relationships and similarities between visual and textual elements. When text or an image/video is mapped to this shared space, they are expected to have similar representations when they convey similar semantics. Once the joint embedding space is learned, tasks such as image captioning, retrieval, or classification can be performed without additional task-specific training. Figure 2 presents the zero-shot action classification inference within this space. For each query video that we want to classify, one can simply look at the closest class label representation to the query video representation. In both Figure 1 and Figure 2, $V_i . T_i$ represents the dot product but is generally a distance or similarity metric between visual and textual embeddings. T_3 here represents the embedding of the ground-truth class label "billiard." $V_i . T_i$ is calculated for all class labels. $V_3 . T_3$ gives the minimum distance with the query video and is therefore colored differently.

2.3. Knowledge distillation

Deep learning models have the capacity to transfer knowledge. For example, after training with the ImageNet dataset, which contains millions of images, a weight file is generated as a result for image classification tasks. Using ImageNet pretrained weights has been proven to be very effective not only for image classification but

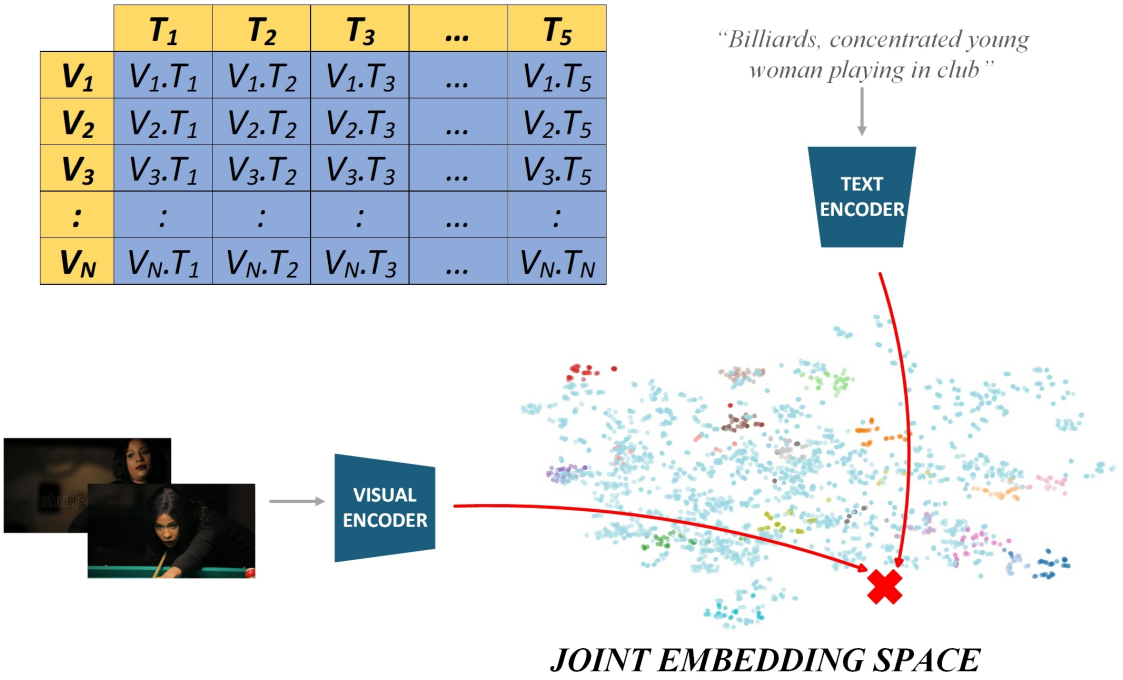


Figure 1. Joint embedding space representation of vision-language model: dual encoder-based architecture and aligned image-text representations.

.

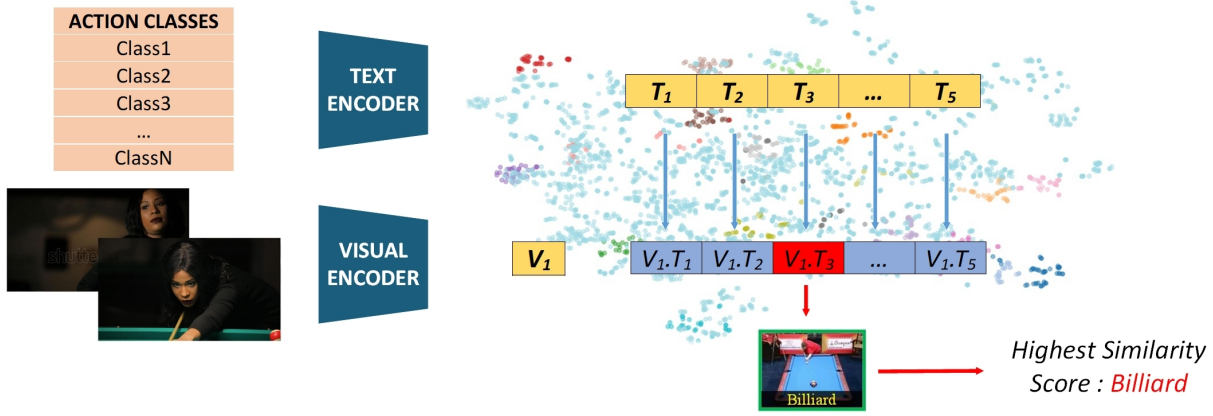


Figure 2. Zero-shot action classification in joint embedding space of vision-language model: nearest-neighbor search among text embeddings of class labels.

also in various computer vision tasks. However, knowledge transfer in 3D CNNs and specifically for action recognition is trickier due to the lack of large-scale datasets for effective pretraining. With the introduction of the Kinetics dataset, the pretraining of networks has yielded significant gains in accuracy [1]. Furthermore, the work in [26] proposed an inflation technique from image models that is also suboptimal due to the risk of adding bias from static image models and constraints with the flexibility of architecture choice. The concept of knowledge distillation was introduced by Hinton [27]. The term typically refers to training a relatively

smaller student network with the same data teacher model trained to achieve the same performance. In this work, we use knowledge distillation for pretraining purposes, similar to DistInit [30]. DistInit uses image-based teachers to generate soft labels for a video model to leverage pretrained 2D networks for learning better video representations. However, DistInit was only studied in the supervised learning context and it used a large set of unlabeled source videos. Specifically, before supervised training with the target dataset, the network is first trained to output the same soft labels as the teacher network, resulting in better weight initialization prior to supervised target task training.

3. Method

Our proposed method is based on feature distillation from a visual-language model. We also take advantage of the strong fine-tuning capability of the visual-language model by using it in the pseudo-labeling of unlabeled samples. Background information on knowledge distillation and visual language models was provided in Sections 2.3 and 2.4, respectively. In this work, we propose a teacher-student training scheme. Our proposed algorithm trains a 3D ResNet-18 from scratch for video action classification in a semisupervised fashion using both labeled and unlabeled videos. The training scheme consists of a feature distillation stage and a fine-tuning stage, which are explained in Sections 3.1 and 3.2.

3.1. Feature distillation stage

During knowledge distillation training, our 3D-CNN is designed to produce feature output matching size with teacher backbone output. We used the space-time encoder proposed in [16] as a multimodal backbone, which utilizes transformer encoders for image, video, and caption encoding. The mentioned network has been trained on the WebVid-2M dataset [16] consisting of videos with captions collected from the internet. We leverage only the visual branch of the mentioned backbone with pretrained weights. We then perform knowledge distillation by enforcing feature consistency. Let $Z = \{z_1, z_2, \ldots, z_K\}$ denote the unlabeled video clips with no class label annotation available. For unlabeled data, given that $g_{student}(x)$ is the visual embedding produced by 3D classifier CNN and $g_{multimodal-teacher}(x)$ is the output visual embedding produced by the frozen teacher network, $L_{distillation}$ given in Eq. (1) is the mean square error (MSE) loss between the multimodal teacher network and classifier to be trained:

$$L_{distillation} = \sum_{Z} ||| g_{student}(x) - g_{multimodal-teacher}(x) ||^2 \tag{1}$$

The concept of knowledge distillation typically entails training a relatively smaller student network with the same data that the teacher model was trained on to achieve the same performance. Our source of unlabeled data is relatively small and different from the one that is used to train the multimodal encoder. We used training data from the UCF101, HMDB51, and Kinetics400 datasets. See Section 4.1 for details about the datasets used. We did not use any labels at this stage. An overview of the feature distillation is depicted in Figure 3. As shown in the figure, in this stage only training videos without labels are presented to the student and teacher networks. The training objective is to minimize the distance between representations produced by the teacher and student. The weights of the teacher network are kept fixed in this stage, preventing any gradient flow for the teacher branch.

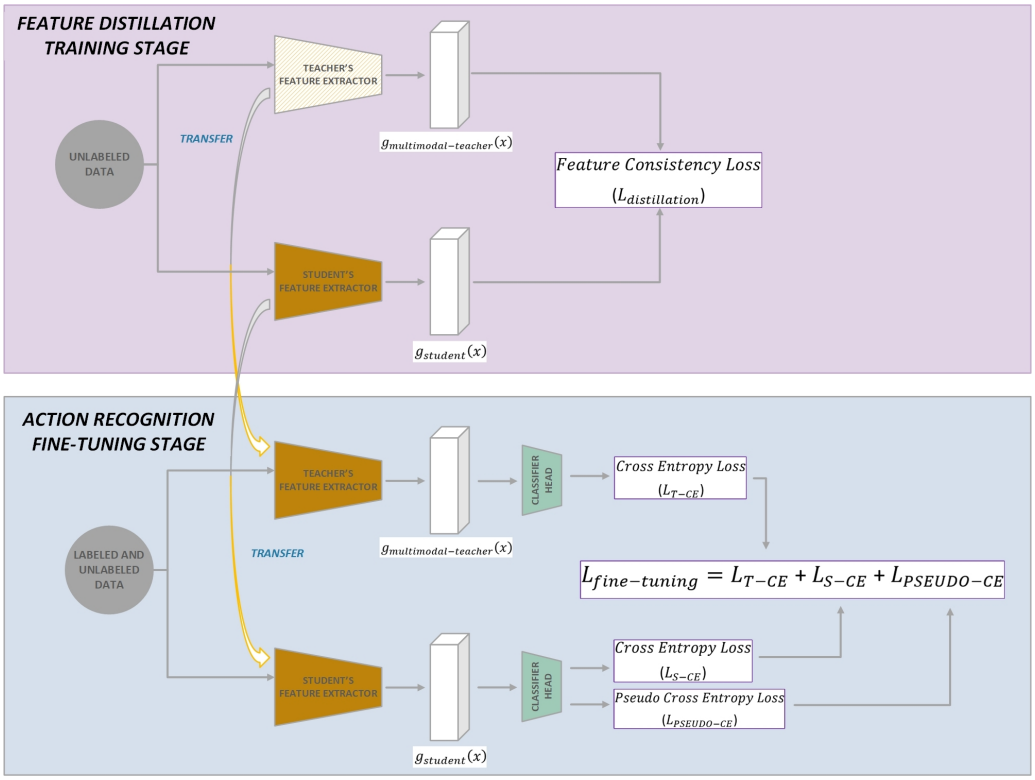


Figure 3. Feature distillation training stage.

3.2. Action recognition fine-tuning stage

In the fine-tuning stage, both the teacher and student networks are fine-tuned for the downstream task of action classification. To do this, we make a small change to the model architectures and add linear classifiers on top of the networks. Let $X_l = \{x_{l_1}, x_{l_2}, \ldots, x_{l_K}\}$ denote the annotated video clips with corresponding label annotations $y_i = \{y_1, y_2, \ldots, y_K\}$. Given C video classes, i.e., $y_i \in \{0,1\}^C$, and given a video and class label pair $P = \{x_{l_i}, y_i\}$ the teacher and student networks are optimized by minimizing cross-entropy losses as given in Eq. (2):

$$L_{\text{T-CE}} = L_{\text{S-CE}} = -\sum_{x_{l_i} \in X} \sum_{c} y_i^c \log p^c(x_{l_i}) \tag{2}$$

After the introduction of labeled data to the teacher network, we assign pseudo-labels to unlabeled data and treat them as "ground-truth" for student training if they are above a certain confidence threshold. The Softmax output of the linear classifier is used as a confidence metric. Let $X_{ul} = x_{ul_1}, x_{ul_2}, \ldots, x_{ul_K}$ denote the unlabeled video clips with corresponding pseudo-label annotations $y_i = y_1, y_2, \ldots, y_K$. Given C video classes, i.e., $y_i \in 0, 1^C$, and a video-label pair $P = x_{ul_i}, y_{ul_i}$, the network is optimized by minimizing the pseudo-cross-entropy loss given in Eq. (3):

$$L_{\text{Pseudo-CE}} = -\sum_{x_{l_i} \in X} \sum_{c} \hat{y_i}^c \log p^c(x_{l_i}) \tag{3}$$

Figure 4. Example pairs of videos and captions from the WebVid2M dataset.

During the fine-tuning stage, the student network is trained using both the calculated cross-entropy loss and the pseudo-cross-entropy loss. For labeled samples in a batch, the cross-entropy loss is calculated, while for unlabeled samples, the pseudo-cross-entropy loss is computed. An overview of the action recognition fine-tuning process is presented in Figure 3. As shown in Figure 3, both labeled and unlabeled videos are presented to the student and teacher networks. In this stage, the weights of the teacher network are no longer fixed. By allowing the gradient flow through the teacher network, we fine-tune it for the target task and get more confident and reliable scores, which are used as pseudo-labels for student training. Pseudo-cross-entropy loss is only calculated for the student network. No self-training of the teacher network is performed.

4. Experiments

4.1. Datasets and evaluation

For video-text pretraining of the teacher multimodal encoder, WebVid-2M was used, which is a large-scale video captioning dataset of over two million pairs of video and text captions. Example video-caption pairs can be seen in Figure 4. For feature distillation, we used the training set of the UCF101 [28] and HMDB51 [29] datasets and an additional 5% of the Kinetics-400 [1] training samples. We did not use any labels in distillation training. Kinetics-400 is a large-scale benchmark dataset for action recognition mostly consisting of YouTube videos. UCF101 and HMDB51, on the other hand, are popular and relatively small-scale benchmark datasets for action recognition models. For performance evaluation, we used the validation set of the UCF101 and HMDB51 datasets. We used the top-1 accuracy metric for performance evaluation.

4.2. Experimental results

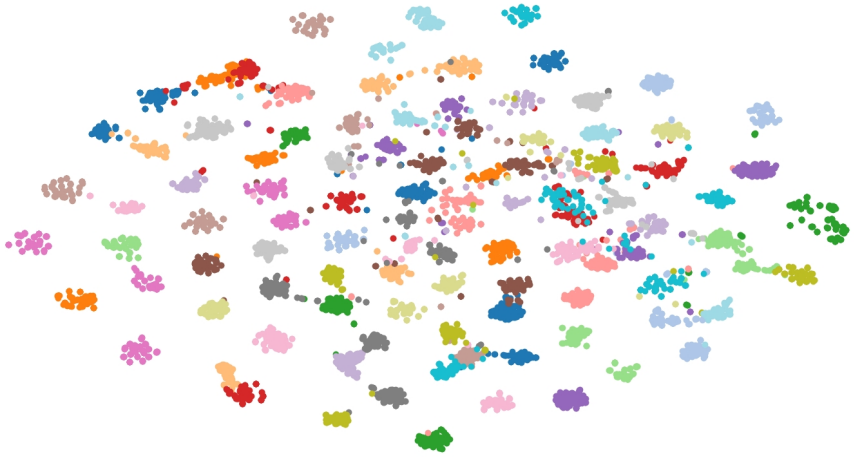
Table 1 shows the zero-shot action classification (ZSAC) performance of the teacher multimodal encoder for the HMDB51 and UCF101 datasets. We evaluate the performance for the official Test-Split-1 for both datasets. For zero-shot inference we use cosine distance as a similarity metric. We assign videos to classes with maximum cosine similarity calculated against class label text embeddings. Figure 5 and Figure 6 show T-SNE visualizations of the UCF101 and HMDB51 datasets, respectively. As shown in Figures 5 and 6, for both datasets, even without fine-tuning, our multimodal encoder can extract representations that cluster videos with respect to action classes, which shows its zero-shot inference capacity. We only fine-tuned the visual encoder of the teacher backbone while keeping the text encoder weights fixed. After fine-tuning, the clusters became even more separable. Table 1 reports top-1 accuracy after fine-tuning with 10% of the labeled data for UCF101 and 50% of the labeled data for HMDB51. The multimodal encoder appears to be a strong fine-tuner and it is fine-tuned quickly after approximately only 10 epochs in our experiments. Our proposed training aims to exploit these learned embeddings. Multimodal learning heavily depends on data. The more data used for training, the larger the architecture will be. Additionally, for applications where architecture customization is required, fine-tuning may not be a practical option. In this work we tried to train a lightweight 3D CNN from scratch using only feature distillation without considering network heterogeneity or any intermediate representation distillation.

Table 1. Top-1 accuracies for multimodal encoder for action classification task.

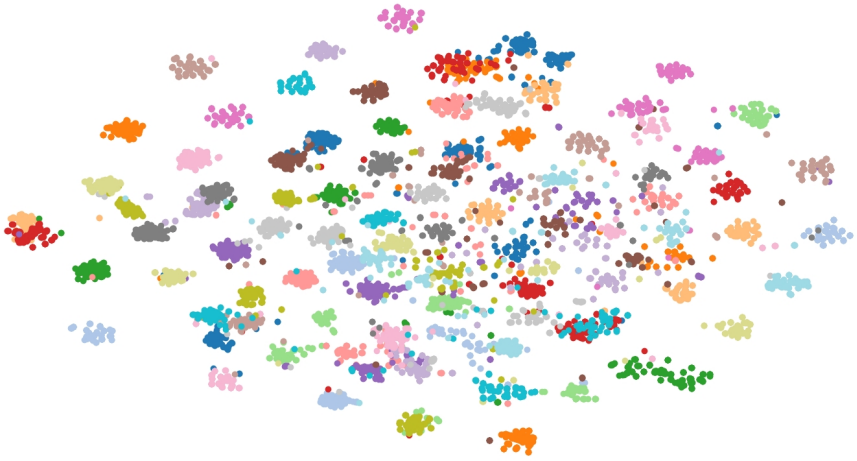
Model	**UCF101 (10%)**	**HMDB51 (50%)**
Teacher-ZSAC	45.5%	27.0%
Teacher-FineTuned	82.5%	57.4%
3D-ResNet-18 (from scracth)	21.6%	29.1%

Table 2 shows a comparison with the state-of-the-art. In Table 2, bold values represent the best values. The compared literature results in this table are derived from [24]. As mentioned in Section 2.2, one avenue of study in the semisupervised action recognition literature proposes the usage of regulatory signals from fixed/frozen pretrained networks for knowledge distillation. We first compare our proposed method with this line of work since we also propose a distillation-based method. DANet and VideoSSL leverage knowledge distillation in the form of feature distillation and logit distillation, respectively. Both DAnet and VideoSSL adopt 2D teacher networks exclusively for knowledge distillation, while DANet employs a combination of three distinct 2D teacher networks. However, recent SOTA methods in computer vision tasks rely on foundation models for representation learning. In contrast to previous work, we propose distilling from a large-scale text-guided pretrained model.

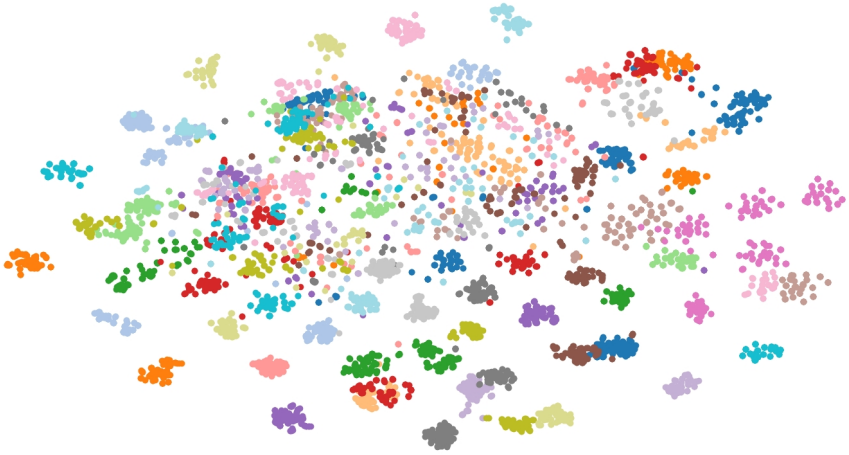
SOTA non-distillation-based methods perform some form of FixMatch [7] like augmentation for the pseudo-labeling process, which requires the processing of the same data when different kinds and levels of augmentations such as horizontal flip, RandAugment, or CTAugment are applied. Although this does not require additional costs for inference, it makes training more complex, especially when new modalities are introduced like temporal gradients [24]. In the official implementation of [24], precalculated and extracted temporal gradient frames from video clips are presented to the model, which requires a decent amount of memory even for relatively small datasets like UCF101 and HMDB51. CMPL [23] applies strong and weak augmentations in two parallel primary and auxiliary backbones with different depths. The mentioned backbones are then asked to predict pseudo-labels for each other. This FixMatch-like auxiliary network requires more frames in training. These works reported 79.1% and 25.1% top-1 accuracies for UCF101 using 10% and 1% of the labeled data, respectively, with a 3D-ResNet-50 primary backbone. For a fair comparison, we only included 3D-ResNet-18



(a) Video-side tuned multimodal encoder features

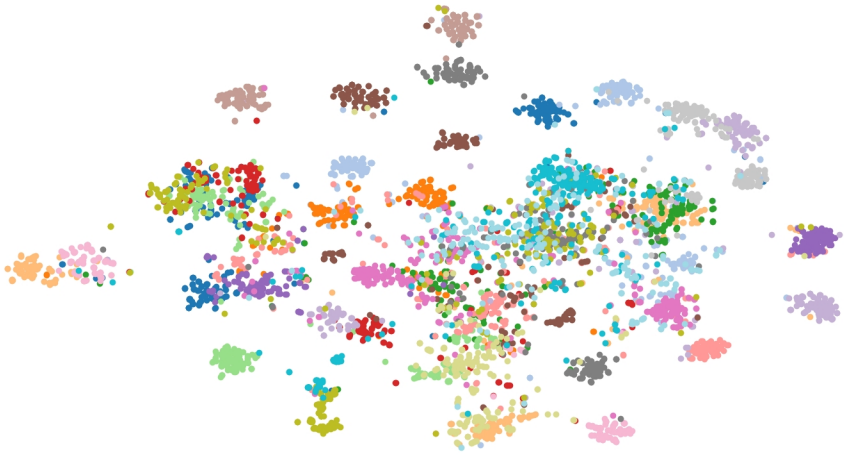


(b) Student network features



(c) Multimodal encoder features before fine-tuning

Figure 5. T-SNE visualizations of extracted features for UCF101 dataset.



(a) Video-side tuned multimodal encoder features

(b) Student network features

(c) Multimodal encoder features before fine-tuning

Figure 6. T-SNE visualizations of extracted features for HMDB51 dataset.

Table 2. Performance comparison with SOTA.

Method	Backbone	Input	UCF101 (10%)	UCF101 (5%)	UCF101 (1%)	HMDB51 (50%)
Supervised	3D-ResNet-18	Video	21.6	15.1	6.5	29.1
Pseudo-Label [31]	3D-ResNet-18	Video	24.7	17.6	-	32.4
Mean Teacher [9]	3D-ResNet-18	Video	25.6	17.5	-	30.4
S4L [33]	3D-ResNet-18	Video	29.1	22.7	-	31.0
ActorCutMix [34]	R(2+1)D-ResNet-18	Video	40.2	27.0		38.2
VideoSSL [20]	3D-ResNet-18	Video	42.0	30.9	-	34.9
DANet [21]	3D-ResNet-18	Video	64.6	-	-	-
LTG [24]	3D-ResNet-18	Video+TG	62.4	44.8	-	**48.4**
Cross-model [23]	3D-ResNet-18	Video	**67.6**	-	23.8	-
L2A[56]	3D-ResNet-18	Video	56.1	-	-	43.2
TACL[49]	3D-ResNet-18*	Video	55.6	-	-	40.2
Proposed Method	3D-ResNet-18	Video	62.4	**53.6**	**24.2**	34.5

*: Indicates modification.

Table 3. Effect of pseudo-labeling.

Labeled data ratio	Pseudo-labeling	Top-1 accuracy
10%	Yes	62.4 %
	No	50.1%
5%	Yes	53.6%
	No	37.5%
1%	Yes	24.2%
	No	13.0%

backbone results in our table. Our proposed method has higher top-1 accuracy for 5% and 1% labeled cases whereas [23] has higher accuracy for 10% labeled cases for UCF101 for the ResNet-18 backbone. The proposed method has higher accuracy in the low-label regime due to better weight initialization after the distillation stage. We also did not include the work in [41] or [32] since those studies did not use a CNN-type architecture and leveraged two additional modalities, respectively.

We reached 62.4%, 53.6%, and 24.2% top-1 accuracies for the UCF101 dataset using only 10%, 5%, and 1% of labeled samples, respectively, by simply using feature distillation and leveraging the teacher network in the pseudo-labeling process. The HMDB51 dataset is more difficult compared to UCF101. We achieved 34.5% accuracy for HMDB51. As shown in Figures 6b and 7b, the T-SNE visualizations of learned embeddings are well separable. Table 3 shows the effect of pseudo-labeling on the top-1 accuracy performance for UCF101. The teacher-generated pseudo-labels improved the accuracy in all three cases differing in terms of the labeled data ratio. Our results show that we reached a compatible performance regarding SOTA methods using a lightweight architecture by making use of a single parallel network when we exploited the multimodal backbone with a vanilla form of feature distillation accompanied by pseudo-labeling. Although our multimodal teacher was previously trained on 2.5M video-caption pairs, our student could mimic its robust task-agnostic features while training a relatively small amount of data (see Section 4.1 for dataset details). Training with more data would be expected to improve the performance more.

4.3. Implementation details

PyTorch implementation of the 3D ResNet-18 model is used. The model is initialized with the "weights=None" argument and training starts from scratch for the student network. We use the AdamW optimizer starting with

a learning rate of 3×10^{-5} for both the feature distillation stage and the action recognition fine-tuning stage. For the distillation stage, training is performed with a batch size of 32. For the action recognition fine-tuning stage, we initially perform training with only cross-entropy loss. For the action recognition fine-tuning stage, the batch size is 4 initially. Once the teacher network predictions are confident enough, model weights are frozen and the batch size is increased to 16. We train the student network for 300 epochs for knowledge distillation. We randomly select 4 frames for each clip. The output size of the backbone visual encoders is 256. We randomly select labeled examples from each of the datasets. The SoftMax output of the linear classifier is used as a confidence metric and no other hypermeter or temperature scaling is used. The confidence threshold is selected as 0.2 and 0.4 for UCF101 and HMDB51, respectively. The training and testing phases of the proposed method are performed with NVIDIA GEFORCE RTX 2080 TI GPU.

5. Discussion

Video action recognition is a fundamental task of video understanding. Due to the annotation labor of supervised learning, especially for large-scale video datasets, label-efficient learning strategies are required. Existing SSL approaches for action recognition leverage video-specific augmentation strategies, FixMatch-like training schemes, additional modalities like temporal gradients or optical flow, and distillation from models previously trained on 2D still-images. In this work, we proposed a distillation-based SSL framework for video action recognition. Unlike previously proposed distillation-based methods, instead of using pretrained fixed-weight still-image networks, we used a vision-language model trained on pairs of videos and captions. During distillation we trained the network using only training videos without any labels at all. In the fine-tuning phase, we used only part of the labeled training data and used the teacher network to pseudo-label the rest. In this study, we aimed to reduce the dependence of action recognition methods on labeled data through distillation from a vision-language model.

We performed an evaluation of our algorithm based on two of the most widely used datasets. We compared our proposed method against several baseline methods in the literature. The main contribution of our paper is integrating a vision-language model into the SSL framework. Our results demonstrate that the proposed feature distillation pretraining improves the performance of supervised learning. Even though we used a relatively small and novel training set for distillation compared to the vision-language model originally trained, knowledge transfer occurred (see Section 4.1. for details about the datasets used). When coupled with pseudo-labeling, the proposed method achieves competitive performance compared to SOTA methods. The architecture of our proposed method is flexible in terms of the choice of teacher and student networks. Therefore, any other foundation model has the potential to be easily plugged into training. Moreover, leveraging more sources of unlabeled data would be expected to improve the performance.

Although our findings show promising results, different forms of knowledge distillation to deal with network heterogeneity and the effect of different distance metrics as loss functions while enforcing feature consistency should be investigated and optimized for customized settings. The limitations of this work can be summarized as follows:

- Knowledge transfer still heavily depends on data size, although not necessarily labeled data.

- The effects of distribution shifts between text-guided pretraining, distillation, and target datasets on knowledge transfer are yet to be explored.

6. Conclusion

In this work, we have proposed an approach for semisupervised action recognition exploiting multimodal feature extractor backbones. We suggest training our student network to mimic the teacher network initially for knowledge transfer using distillation loss for feature consistency. The mentioned distillation stage provides better weight initialization and this results in higher performance in the case of less labeled data. We also propose fine-tuning the teacher network for pseudo-labeling. Since the multimodal teacher backbone is a strong fine-tuner, we generated pseudo-labels with high accuracy, improving the accuracy further. The experimental results indicated that the proposed method achieves competitive performance compared to state-of-the-art methods. There is a growing body of literature on multimodal learning for generating joint embedding spaces. Our work has aimed to show that vision-language multimodal embedding spaces can be utilized in a semisupervised learning framework for action recognition tasks. We hope that our work will inspire further research on the intersection of multimodal learning and SSL. There is still room for improvement on how to effectively transfer video representations from VL models for label-efficient training. Task-specific fine-tuning of foundation models and the incorporation of video-specific augmentations could further reduce label dependency. In the future, we plan to investigate the effectiveness of feature distillation for other downstream video applications.

Acknowledgment

This work was supported by the Kocaeli University Scientific Research Projects Coordination Unit under grant number FDK-2022-2989.

References

[1] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the Kinetics dataset. In: Conference on Computer Vision and Pattern Recognition; Honolulu, HI, USA; 2017. pp. 6299-6308.

[2] Fan Q, Chen CFR, Kuehne H, Pistoia M, Cox D. More is less: learning efficient video representations by big-little network and depthwise temporal aggregation. In: Conference on Neural Information Processing Systems; Vancouver, BC, Canada; 2019. pp. 2264-2273.

[3] Feichtenhofer C. X3D: Expanding architectures for efficient video recognition. In: Conference on Computer Vision and Pattern Recognition; Seattle, WA, USA; 2020. pp. 203-213.

[4] Feichtenhofer C, Fan H, Malik J, He K. SlowFast networks for video eecognition. In: International Conference on Computer Vision; Seoul, Korea; 2019. pp. 6202-6211.

[5] Lin J, Gan C, Han S. TSM: Temporal shift module for efficient video understanding. In: International Conference on Computer Vision; Seoul, Korea; 2019. pp. 7083–7093.

[6] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: International Conference on Computer Vision; Santiago, Chile; 2015. pp. 4489-4497.

[7] Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: Conference on Neural Information Processing Systems; Vancouver, BC, Canada; 2020. pp. 596-608.

[8] Pham H, Dai Z, Xie Q, Le QV. Meta pseudo labels. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Nashville, TN, USA; 2021. pp. 11557-11568.

[9] Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Conference on Neural Information Processing Systems; Long Beach, CA, USA; 2017. pp. 1195-1204.

[10] Berthelot D, Carlini N, Cubuk ED, Kurakin A, Sohn K et al. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv preprint. arXiv:1911.09785. 2019.

[11] Frome A, Corrado GS, Shlens J, Bengio S, Dean J et al. Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems; Lake Tahoe, NV, United States; 2013. pp. 2121-2129.

[12] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G et al. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning; Virtual; 2021. pp. 8748-8763.

[13] Xu X, Hospedales T, Gong S. Transductive zero-shot action recognition by word-vector embedding. International Journal of Computer Vision 2017; 123 (3): 309-333. https://doi.org/10.1007/s11263-016-0983-5

[14] Chen B, Rouditchenko A, Duarte K, Kuehne H, Thomas S et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. In: International Conference on Computer Vision; Virtual; 2021. pp. 8012-8021.

[15] Brattoli B, Tighe J, Zhdanov F, Perona P, Chalupka K. Rethinking zero-shot video classification: end-to-end training for realistic applications. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Seattle, WA, USA. 2020. pp. 4613-4623.

[16] Bain M, Nagrani A, Varol G, Zisserman A. Frozen in time: a joint video and image encoder for end-to-end retrieval. In: International Conference on Computer Vision; Virtual; 2021. pp. 1728-1738.

[17] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems; Montreal, QC, Canada; 2014. pp. 1-9.

[18] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X et al. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations; Virtual; 2021. pp. 1-21.

[19] Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. In: International Conference on Computer Vision Workshops; Venice, Italy; 2017. pp. 3154-3160.

[20] Jing L, Parag T, Wu Z, Tian Y, Wang. VideoSSL: Semi-supervised learning for video classification. In: Winter Conference on Applications of Computer Vision; Virtual; 2021. pp. 1110-1119.

[21] Gao G, Liu Z, Zhang G, Li J, Qin AK. DANet: Semi-supervised differentiated auxiliaries guided network for video action recognition. Neural Networks 2023; 158: 121-131. https://doi.org/10.1016/j.neunet.2022.11.009

[22] Singh A, Chakraborty O, Varshney A, Panda R, Feris R et al. Semi-supervised action recognition with temporal contrastive learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Virtual; 2021. pp. 10389-10399.

[23] Xu Y, We F, Sun X, Yang C, Shen Y et al. Cross-model pseudo-labeling for semi-supervised action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; New Orleans, LA, USA; 2022. pp. 2959-2968.

[24] Xiao J, Jing L, Zhang L, He J, She Q et al. Learning from temporal gradient for semi-supervised action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; New Orleans, LA, USA; 2022. pp. 3252-3262.

[25] Socher R, Ganjoo M, Manning CD, Ng A. Zero-shot learning through cross-modal transfer. In: International Conference on Neural Information Processing System; Lake Tahoe, NV, USA; 2013. pp. 935-943.

[26] Feichtenhofer C, Pinz A, Wildes RP. Spatiotemporal residual networks for video action recognition. arXiv preprint. arXiv:1611.02155. 2016.

[27] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint. arXiv:1503.02531. 2015.

[28] Soomro K, Zamir AR, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint. arXiv:1212.0402. 2012.

[29] Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB: A large video database for human motion recognition. In: International Conference on Computer Vision; Barcelona, Spain; 2011. pp. 2556-2563.

[30] Girdhar R, Tran D, Torresani L, Ramanan D. Distinit: Learning video representations without a single labeled video. In: International Conference on Computer Vision; Seoul, Korea; 2019. pp. 852-861.

[31] Lee DH. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks. In: ICML 2013 Workshop: Challenges in Representation Learning; Atlanta, GA, USA; 2013. pp. 1-6.

[32] Xiong B, Fan H, Grauman, Feichtenhofer C. Multiview pseudo-labeling for semi-supervised learning from video. In: International Conference on Computer Vision; Montreal, QC, Canada; 2021. pp. 7209-7219.

[33] Zhai X, Oliver A, Kolesnikov A, Beyer L. S4l: Self-supervised semi-supervised learning. In: International Conference on Computer Vision; Seoul, Korea; 2019. pp. 1476-1485.

[34] Zou Y, Choi J, Wang Q, Huang J. Learning representational invariances for data-efficient action recognition. arXiv preprint. arXiv:2103.16565. 2021.

[35] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R et al. Large-scale video classification with convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Columbus, OH, USA; 2014. pp. 1725-1732.

[36] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Las Vegas, NV, USA; 2016. pp. 1933-1941.

[37] Lei J, Li L, Zhou L, Gan Z, Berg TL et al. Less is more: Clipbert for video-and-language learning via sparse sampling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Nashville, TN, USA; 2021. pp. 7331-7341.

[38] Arnab A, Dehghani M, Heigold G, Sun C, Lučić M. Vivit: A video vision transformer. In: International Conference on Computer Vision; Montreal, QC, Canada; 2021. pp. 6836-6846.

[39] Liu Z, Ning J, Cao Y, Wei Y, Zhang Z. Video swin transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; New Orleans, LA, USA 2022. pp. 3202-3211.

[40] Yan S, Xiong X, Arnab A, Lu Z, Zhang M et al. Multiview transformers for video recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; New Orleans, LA, USA; 2022. pp. 3333-3343.

[41] Xing Z, Dai Q, Hu H, Che J, Wu Z et al. Svformer: Semi-supervised video transformer for action recognition. In: Conference on Computer Vision and Pattern Recognition; Vancouver, BC, Canada; 2023. pp. 18816-18826.

[42] Miech A, Zhukov D, Alayrac JB, Tapaswi M, Laptev. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: IEEE/CVF International Conference on Computer Vision; Seoul, Korea; 2019. pp. 2630-2640.

[43] Zellers R, Lu X, Hessel J, Yu Y, Park JS. Merlot: Multimodal neural script knowledge models. In: 35th Conference on Neural Information Processing Systems; Virtual; 2021. pp. 23634-23651.

[44] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Salt Lake City, UT, USA; 2018. pp. 6546-6555.

[45] Feichtenhofer C. X3d: Expanding architectures for efficient video recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Seattle, WA, USA; 2020. pp. 203-213.

[46] Xie S, Sun C, Huang J, Tu Z, Murphy K. Rethinking spatiotemporal feature learning for video understanding. arXiv preprint. arXiv:1712.04851. 2017.

[47] Liu Z, Luo D, Wang Y, Wang L, Tai Y et al. Teinet: Towards an efficient architecture for video recognition. In: AAAI Conference on Artificial Intelligence; New York, NY, USA; 2020. pp. 11669-11676.

[48] Li Y, Ji B, Shi X, Zhang J, Kang B. Tea: Temporal excitation and aggregation for action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Seattle, WA, USA; 2020. pp. 909-918.

[49] Tong A, Tang C, Wang W. Semi-supervised action recognition from temporal augmentation using curriculum learning. IEEE Transactions on Circuits and Systems for Video Technology 2022; 33 (3): 1305-1319. https://doi.org/10.1109/TCSVT.2022.3210271

[50] Liu Z, Wang L, Wu W, Qian C, Lu T. TAM: Temporal adaptive module for video recognition. In: International Conference on Computer Vision; Montreal, QC, Canada; 2021. pp. 13708-13718.

[51] Wang L, Tong Z, Ji B, Wu G. TDN: Temporal difference networks for efficient action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Nashville, TN, USA; 2021. pp. 1895-1904.

[52] Neimark D, Bar O, Zohar M, Asselmann D. Video transformer network. In: International Conference on Computer Vision; Montreal, BC, Canada; 2021. pp. 3163-3172.

[53] Kolesnikov A, Zhai X, Beyer L. Revisiting self-supervised visual representation learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Long Beach, CA, USA; 2019. pp. 1920-1929.

[54] Choi J, Gao C, Messou JC, Huang JB. Why can't I dance in the mall? Learning to mitigate scene bias in action recognition. In: 33rd Conference on Neural Information Processing Systems; Vancouver, BC, Canada; 2019. pp. 1-13.

[55] Xie Q, Dai Z, Hovy E, Luong T, Le Q. Unsupervised data augmentation for consistency training. In: 34th Conference on Neural Information Processing Systems; Vancouver, BC, Canada; 2020. pp. 6256-6268.

[56] Gowda SN, Rohrbach M, Keller F, Sevilla-Lara L. Learn2augment: Learning to composite videos for data augmentation in action recognition. In: European Conference on Computer Vision; Tel Aviv, Israel; 2022. pp. 242-259.

[57] Jia C, Yang Y, Xia Y, Chen YT, Parekh Z et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning; Virtual; 2021. pp. 4904-4916.

[58] Yuan L, Chen D, Chen YL, Codella N, Dai X et al. Florence: A new foundation model for computer vision. arXiv preprint. arXiv:2111.11432. 2021.

[59] Fu TJ, Li L, Gan Z, Lin K, Wang WY. Violet: End-to-end video-language transformers with masked visual-token modeling. arXiv preprint. arXiv:2111.12681. 2021.