creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Master's Thesis

Semi-Supervised Multi-modal Video Action Recognition with Audio Source Localization Guided Mixup

Seok Un Kang

Graduate School of Artificial Intelligence
(Artificial Intelligence)

Ulsan National Institute of Science and Technology

2024

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Semi-Supervised Multi-modal Video Action Recognition with Audio Source Localization Guided Mixup

Seok Un Kang

Graduate School of Artificial Intelligence
(Artificial Intelligence)

Ulsan National Institute of Science and Technology

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Semi-Supervised Multi-modal Video Action Recognition with Audio Source Localization Guided Mixup

A thesis/dissertation submitted to
Ulsan National Institute of Science and Technology
in partial fulfillment of the
requirements for the degree of
Master of Science

Seok Un Kang

01.19.2024 of submission
Approved by

Advisor
Tae Hwan Kim

Semi-Supervised Multi-modal Video Action Recognition with Audio Source Localization Guided Mixup

Seok Un Kang

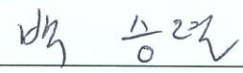
This certifies that the thesis/dissertation of Seok Un Kang is approved.

01.19.2024 of submission

Signature

Advisor: Tae Hwan Kim

Signature



Seung Ryul Baek

Signature

Hyoung Hun Kim

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Abstract

Video action recognition is a challenging but important task to understand and find out what the video does. However, acquiring labels of video is costly, and semi-supervised learning (SSL) has been studied to improve the performance even with the small number of labeled data in the task. Prior studies for semi-supervised video action recognition have mostly focused on using single modality - visuals - but video is multi-modal so utilizing both visuals and audio would be desirable and improve the performance further, which has not been well explored. Therefore, we propose audio-visual SSL for video action recognition, which uses both visual and audio together, even with quite a few labeled data that is challenging. In addition, to maximize the information of audio and video, we propose a novel audio source localization-guided mixup method that considers inter-modal relations between video and audio modalities. In experiments on UCF-51, Kinetics-400, and VGGSound datasets, our model shows the superior performance of the proposed SSL audio-visual action recognition and audio source localization-guided mixup.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Contents

I Introduction 1

II Related Works 3

2.1 Semi-supervised Image Classification 3

2.2 Semi-supervised Video Recognition 3

2.3 Audio Source Localization 4

III Methodology 5

3.1 Preliminaries 5

3.2 Model Pipeline 7

3.3 Audio Source Localization-guided Mixup 8

3.4 Visual-Audio Contrastive Learning 10

3.5 Training Objective 11

IV Experiment 12

4.1 Experimental Setting 12

4.2 Implementation Details 12

4.3 Main Results 13

4.4 Ablation Studies 14

V Conclusion 16

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

References 17

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

List of Figures

1 **Overview of our audio-visual semi-supervised video action recognition framework.** (a) For video clip A, an Audio Source Localization Model generates the audio source localization map (ASL Map), which is used to create a new mask for inter-modal relations. Audio mixup is achieved through pixel-level interpolation of Log Mel-filter bank coefficients. In the teacher model, predictions from an unmixed video clip input are interpolated by the mask ratio λ to form pseudo labels for the consistency regularization loss L_{mix} in unlabeled data mixtures. This model is continuously updated via the student model's exponential moving average. (b) When performing visual-audio contrastive learning, labeled, unlabeled, and mixed data are all concatenated and used together for calculating the similarities. 5

2 **Overview of our Audio Source Localization-guided Mixup Framework.** The framework performs the audio source localization-guided mixup on video clips A and B. For generating the audio source localization map, video clip A is utilized. The video and audio from clip A are processed through an audio source localization model to produce the localization maps. This generated map is then used as the weight for performing multinomial sampling without replacement, creating an audio source localization-guided mask. This mask guides semantically important regions in video A, taking into account the audio information. Consequently, our proposed audio source localization-guided mixup allows consideration of the interrelation between video and audio modalities sharing the same video clip. For audio A and B, log mel-filterbank coefficients are transformed and interpolated at the pixel level. The resulting mixed video and audio are then used as input for prediction. 6

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

3 **Visualization of the TubeToken mask and our proposed Audio Source Localization-guided Mask.** When an original image (a) is provided, applying the TubeToken masking generates a mask with a random pattern (b). The created TubeToken mask is then utilized to produce a masked image, as shown in (c), which is used as input in SVFormer. Sampling this TubeToken mask 255 times allows for the observation of a random average TubeToken Mask, as seen in (d). In contrast, our proposed method involves creating the audio source localization-guided mask. This process starts with the original input (a), from which a localization map is generated, as shown in (e). This map is then employed as a weight for sampling, leading to the generation of the audio source localization-guided mask (f). Applying this mask to the original input results in a modified input, as depicted in (g). Furthermore, similar to the TubeToken mask, sampling the mask 255 times produces a result like (h). This process enables the capture of areas where the audio source is located, allowing us to consider the interrelation between the visual and audio modalities and perform the mixup accordingly. 9

4 **Performance Impact of Hyper-parameter τ.** This figure presents the results of experiments conducted to explore the performance impact of the hyper-parameter τ. It specifically focuses on the changes in performance with varying thresholds of τ during training with only one labeled sample in the UCF-51 dataset. 15

5 **Ablation study on the effect of varying frame counts (1, 2, 4, and 8 frames) on the audio source localization map.** This study evaluates the impact of averaging the localization maps over different numbers of frames before using them in the sampling process. 16

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

I Introduction

Video is considered an important resource for deep learning vision research. Video data of different lengths and formats has become more readily available as large-scale video content offerings and user interactions on online platforms increase. This leads to active research on video understanding [1–4] and among them, video action recognition is one of the challenging tasks [5–7].

Unlike image classification, video action recognition is a challenging task that requires understanding not only spatial but also temporal information. This requires sufficient labeled training data, but labeling on video data is more difficult and time-consuming than images. Nevertheless, unlabeled video data is readily available, and semi-supervised learning (SSL) methods using labeled video and large-scale unlabeled video together are being actively studied [8–11].

Semi-supervised video action recognition research remains less explored compared to semi-supervised image classification [12–16]. Recent studies in SSL video recognition have actively explored various approaches beyond just using fixed, flexible, or even not predefined confidence-based thresholds [12,13,15] as proposed in the SSL image domain. These include leveraging pre-trained networks and large-scale unlabeled datasets [9, 17, 18]. In addition, research efforts are increasingly focusing on the utilization of additional modalities such as temporal gradient [10] and optical flow [19] obtained from video data or by introducing auxiliary networks [11].

Studies of semi-supervised learning for image and video use random augmentation techniques to generate weak and strong augmented images or videos, and train the model through consistency regularization using both augmented [9, 11–13, 15]. In this context, weak augmentations typically involve random flips and rotations, while strong augmentations utilize RandAugment [20] that includes techniques such as CutMix [21], which involves cutting and combining different images, and MixUp [22], which mixes images through interpolation.

However, SVFormer [9] points out the limitations of mixup and cutmix that do not produce the expected performance in the video domain, and to overcome these limitations, they propose TubeToken Mixup and Temporary Warping augmentation methods. However, there is still a limitation that they cannot actively utilize all its information because only visual information is considered, but not audio information. In semi-supervised video action recognition, using visual-audio modality has been underexplored. The only exception is AvCLR [23], which employs audio-visual contrastive learning with ResNet architecture. However, despite originating from the same video clip, augmentations for video and audio modalities are conducted individually, without deeply considering the inter-modal relation between these modalities.

In this paper, we propose ASGM-Former, the transformer-based framework for semi-supervised audio-visual action recognition with audio source localization-guided mixup. Specifically, we introduce a novel audio source localization-guided mixup method, designed to preserve the inter-modal relations that tend to be overlooked when applying mixup or cutmix to video and audio modality. In addition, we leverage contrastive learning between video and audio modalities in the video clip. Incorporating audio source localization guided mixup with contrastive learning can achieve substantial performance

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

improvements over existing state-of-the-art methods.

In summary, our contributions are as follows:

- We study semi-supervised audio-visual action recognition which has been under-explored. To the best of our knowledge, we propose, for the first time, a visual-audio semi-supervised video action recognition approach based on the transformer model. This allows for the maximization of the use of visual and audio information inherent in videos.
- We propose a novel audio source localization-guided mixup that, unlike mixup or cutmix, preserves the interrelation between the video and audio modalities.
- In experiments, our proposed model outperforms the existing state-of-the-art model on three different benchmark datasets in even challenging scenarios with only a few labeled samples available.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

II Related Works

2.1 Semi-supervised Image Classification

Deep learning-based image classification has achieved significant performance improvements by leveraging large-scale annotated datasets [24–29]. However, annotating large-scale image datasets requires considerable time and resources. To address this, semi-supervised learning (SSL) methods have been introduced, which use a few labeled datasets and large-scale unlabeled datasets. These approaches aim to reduce the time and resource consumption for annotations while improving performance. The most commonly used techniques include consistency learning through data augmentations [30, 31] and the use of pseudo labels [32, 33]. Consistency learning focuses on training the model to make consistent predictions across various augmented versions of a single image. In contrast, pseudo-labeling, for example using a teacher model or the model itself, involves making predictions on unlabeled data and using the predicted probabilities as pseudo-labels. Additionally, various studies use consistency learning and pseudo-labeling together to reduce the bias of pseudo-labels according to certain thresholds. For example, FixMatch [12] uses fixed predefined thresholds, FlexMatch [13] employs flexible thresholds, and FreeMatch [15] uses thresholds that are not predefined but are also flexible.

2.2 Semi-supervised Video Recognition

As the creation and utilization of video data increase across various platforms, research on video understanding is becoming increasingly active [1–4, 17]. Among them, video recognition [6, 34–38] is one of the challenging tasks. Similar to image classification, video recognition requires a large amount of labeled data. However, videos have more complex information than images, which needs significant time and resources for annotations. To address this challenge, various studies have been conducted to extend semi-supervised learning techniques used in image classification to video recognition [8, 11, 39, 40].

VideoSSL [18] extends the existing semi-supervised image classification methods to the video modality. In addition, MvPL [19] has introduced a multi-modal approach that leverages optical flows that represent motion by detailing the instant velocities of images across the horizontal and vertical axes, and temporal gradients by calculating temporal differences between two consecutive frames. On the other hand, SVFormer [9] recognizes the limitation that image-based augmentation methods, such as mixup, do not perform well in the video domain. To overcome this limitation, SVFormer introduces TubeToken Mixup which is mixing the two different videos at the token level feature. In SSL video recognition, studies considering video and audio modalities are scarce, with AvCLR [23] being a notable exception. However, its augmentation methods focus on individual modalities, overlooking the inter-modal relation between video and audio. Therefore, We introduce a novel approach that employs audio source localization-guided mixup to bridge this gap, considering the inter-modal relation between video and audio modality.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

2.3 Audio Source Localization

Audio source localization is the field of research that combines visual and audio information to identify the source of audio within the scene [41–43]. Senocak et al. [41] propose the two-stream network architecture that uses attention mechanisms for visual and audio modalities through unsupervised learning. Hu et al. [44] introduce multi-source audio-visual sound localization utilizing contrastive learning and cycle consistency using the synthetic mixture of audio from multiple videos. EZ-VSL [45] proposes multiple instance contrastive learning for alignment between audio-visual modalities and the object-guided localization scheme that utilizes an object localization map generated from a pre-trained visual encoder. FNAC [46] proposes False Negatives Suppression that uses intra-modal similarity to identify potential false negatives and minimize their negative impact, and True Negatives Suppression that encourages more discriminatory sound source localization by highlighting true negatives using different localization results. We propose an augmentation methodology that utilizes audio source localization to consider the inter-modal relation of video and audio modalities.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

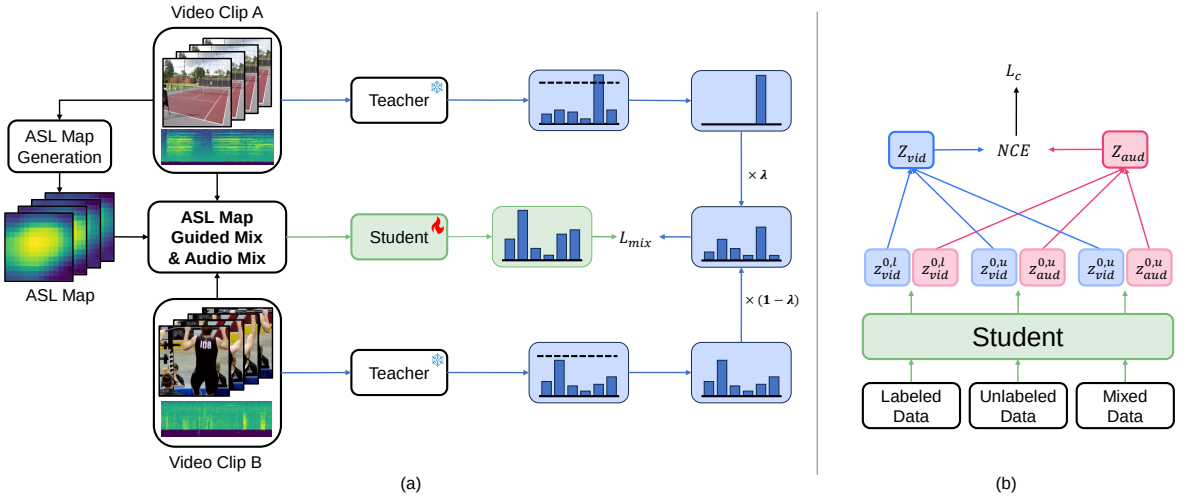


Figure 1: **Overview of our audio-visual semi-supervised video action recognition framework.** (a) For video clip A, an Audio Source Localization Model generates the audio source localization map (ASL Map), which is used to create a new mask for inter-modal relations. Audio mixup is achieved through pixel-level interpolation of Log Mel-filter bank coefficients. In the teacher model, predictions from an unmixed video clip input are interpolated by the mask ratio λ to form pseudo labels for the consistency regularization loss L_{mix} in unlabeled data mixtures. This model is continuously updated via the student model's exponential moving average. (b) When performing visual-audio contrastive learning, labeled, unlabeled, and mixed data are all concatenated and used together for calculating the similarities.

III Methodology

In this section, first, we describe the methodologies primarily used in semi-supervised learning. Then, we introduce the multi-modal video recognition framework that extends to video-audio modalities from the existing video recognition framework. We then introduce our proposed audio source localization-guided mixup methodology.

3.1 Preliminaries

In semi-supervised learning, there are few labeled samples and large-scale unlabeled samples. Let $D_L = \{(x^l, y^l) : l \in [N_L]\}$ as labeled dataset and $D_U = \{(x^u) : u \in [N_U]\}$ as unlabeled dataset where N_L and N_U are numbers of samples in labeled and unlabeled dataset respectively, and in general $N_U \gg N_L$. For the labeled dataset, like the general classification problem, training is conducted through standard cross-entry loss between predicted probabilities and labels like Eq. 1.

$$L_s = \frac{1}{B_l} \sum_{i}^{B_l} H(y_i^l, p_m(x_i^l)) \tag{1}$$

where B_l is batch size of the labeled data and $p_m(\cdot)$ is denoted predicted probability from the model. In the case of the unlabeled data x^u, consistency regularization is used through data augmentation like

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

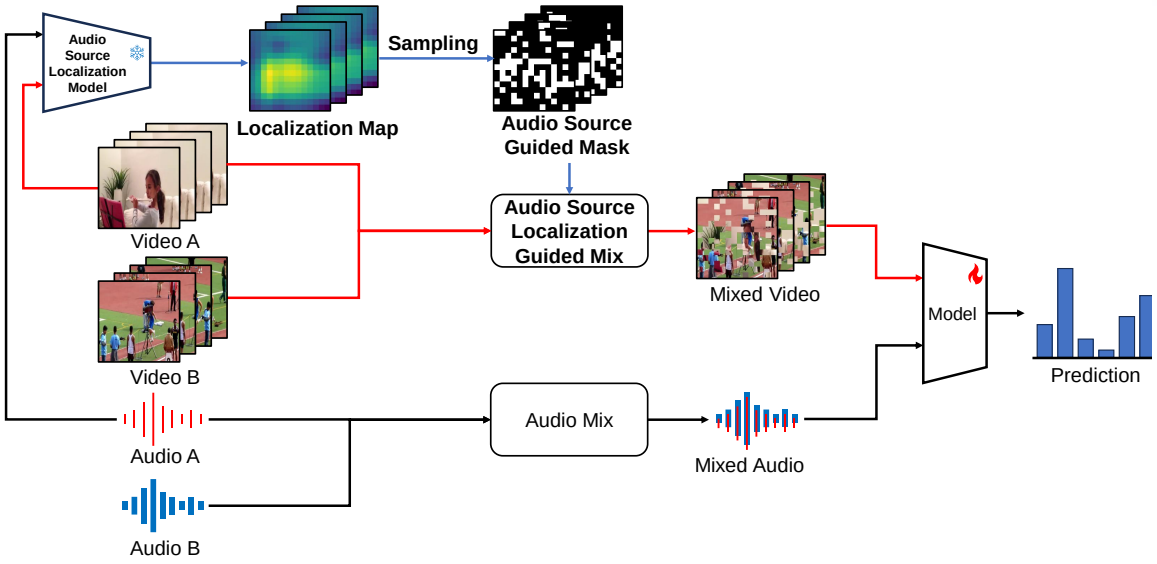


Figure 2: **Overview of our Audio Source Localization-guided Mixup Framework.** The framework performs the audio source localization-guided mixup on video clips A and B. For generating the audio source localization map, video clip A is utilized. The video and audio from clip A are processed through an audio source localization model to produce the localization maps. This generated map is then used as the weight for performing multinomial sampling without replacement, creating an audio source localization-guided mask. This mask guides semantically important regions in video A, taking into account the audio information. Consequently, our proposed audio source localization-guided mixup allows consideration of the interrelation between video and audio modalities sharing the same video clip. For audio A and B, log mel-filterbank coefficients are transformed and interpolated at the pixel level. The resulting mixed video and audio are then used as input for prediction.

weak-augmentation $\alpha_{vid}(\cdot)$ (e.g., random flip or rotation) and strong-augmentation $A_{vid}(\cdot)$ (e.g., RandAugmentation [20]). Throughout these two different levels of augmentation, consistency learning is conducted for unlabeled data like Eq. 2.

$$L_u = \frac{1}{B_u}\sum_{i}^{B_u} H(\hat{p}_m(\alpha_{vid}(x_i^u)), p_m(A_{vid}(x_i^u))) \tag{2}$$

where B_u is the number of unlabeled data in batch and $\hat{p}_m(\cdot)$ can be predicted probability from the Mean Teacher model or model itself by freezing the parameters. And let $\hat{p}_m(\alpha_{vid}(x_i^u))$ as $\hat{q}_i^u$ and $p_m(A_{vid}(x_i^u))$ as Q_i^u. In this case, the predicted probability served as the pseudo label is not reliable so FixMatch [12] proposes using fixed threshold τ for remaining confidential probability as the pseudo label like Eq. 3.

$$L_u = \frac{1}{B_u}\sum_{i}^{B_u} \mathbb{1}(\max(\hat{q}_i^u) \geq \tau) H(\hat{q}_i^u, Q_i^u) \tag{3}$$

where $\mathbb{1}(\cdot)$ is the indicator function.

Unlike the fixed threshold approach which uses a single threshold for all classes, the flexible threshold approach in FlexMatch [13] sets and updates different thresholds for each class. In Eq.3, τ is replaced

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

with $T = [\tau_1, \cdots, \tau_c]$, where c is the number of classes. The flexible threshold for each class is represented by $\tau(\text{argmax}(\hat{q}_i^u))$, where $\text{argmax}(\hat{q}_i^u)$ identifies the class with the highest predicted probability for a given unlabeled sample.

To update the confidence threshold for each class at the t^{th} step, the learning effect for class c can be calculated as shown in Eq 4. Here, τ represents the predefined fixed threshold used in FixMatch. After calculating the learning effect for class c, the confidence threshold for class c is updated as in Equation 5, wherein normalization is applied based on the number of samples surpassing the threshold, about the class that exhibits the maximal count of such samples. The loss function for unlabeled data, as shown in Eq.6.

$$\sigma_t(c) = \sum_{n}^{N_U} \mathbb{1}(\max(\hat{q}_n^u) \geq \tau)\mathbb{1}(\text{argmax}(Q_n^u) = c) \tag{4}$$

$$T(c) = \frac{\sigma_t(c)}{\max_c \sigma_t}\tau \tag{5}$$

$$L_u = \frac{1}{B_u}\sum_{i}^{B_u} \mathbb{1}(\max(\hat{q}_i^u) \geq T(\text{argmax}(\hat{q}_i^u)))H(\hat{q}_i^u, Q_i^u). \tag{6}$$

3.2 Model Pipeline

Our model follows the state-of-the-art semi-supervised learning framework for video action recognition, SVFormer [9]. This framework utilizes the consistency loss based on FixMatch, along with augmentation methods for video modality, namely TubeToken Mix (TTMix) and Temporal Warping Augmentation (TWAug).

Let two unlabeled video clips as $v_a^u, v_b^u \in \mathbb{R}^{H\times W\times C\times T}$ where (H, W) is the resolution of the image frame, C is the number of the channels and T is the number of the frames in the video clip, and let two embedded vectors $Emb_{vid}(v_a^u), Emb_{vid}(v_b^u) \in \mathbb{R}^{N\times(P^2\cdot C)\times T}$ where P is the non-overlapping patch size and $N = HW/P^2$ is the number of the embedding patches. For the mixing of two embedded video clips, token-level mask $\mathbf{M} \in \{0,1\}^{(P^2\cdot C)\times T}$ is utilized as Eq. 7.

$$\begin{aligned} e_{mix}^u = &Emb_{vid}(A_{vid}(v_a^u)) \odot \mathbf{M} \\ &+ Emb_{vid}(A_{vid}(v_b^u)) \odot (1-\mathbf{M}) \end{aligned} \tag{7}$$

where $\odot$ is element-wise multiplication.

At this point, SVFormer utilizes tube-style masking which shares the same mask pattern through all the frames and has consistency in the temporal axis than frame token masking and random token masking. In the case of TWAug, it is utilized for stretching one frame to various temporal lengths. For example, given the number of T frames, they select a few frames and pad with chosen frames at the other frame positions that are not selected. By this TWAug, they argue that the model can learn temporal dynamics flexibly.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

Although most commonly encountered videos have information consisting of sequences of frames and corresponding audio information, many studies pay less attention to audio information. Inspired by this respect, we expand the SVFormer framework to video-audio modalities.

Let unlabeled audio log mel-filterbank feature data as $a_a^u, a_b^u \in \mathbb{R}^{M\times L}$ from corresponding video clips v_a^u, v_b^u, respectively, where L is time-step and M is range of the mel-frequency. For the augmentation, we use SpecAugment [47] as strong-augmentation $A_{aud}(\cdot)$. Through the Eq. 8, mixed audio is conducted.

$$a_{mix}^u = \lambda \cdot A_{aud}(a_a^u) + (1-\lambda)\cdot A_{aud}(a_b^u) \tag{8}$$

where λ is the hyper-parameter that follows the beta-distribution used for the mixup.

The following discussion focuses on the process of learning utilizing video and audio data. Initially, a [cls] token is concatenated to the previously obtained embedding, and this combined embedding is then fed into the video encoder E_{vid}. As a result, the output z_{vid} is calculated. Similarly, for audio, after embedding the audio, it is input into the audio encoder E_{aud}, resulting in the output z_{aud}. This calculation is applied in the same way to video clips v_a^u, v_b^u and their corresponding audio a_a^u, a_b^u as well as to the mixed video and audio obtained from Eq. 7 and 8. This process can be verified through the following equations.

$$z_{vid} = E_{vid}([\text{CLS}_{vid}, e_{mix}^u]) \tag{9}$$

$$z_{aud} = E_{aud}([\text{CLS}_{aud}, Emb_{aud}(e_{mix}^u)]) \tag{10}$$

$$\hat{y} = \text{FUSION}(z_{vid}^0, z_{aud}^0) \tag{11}$$

where z_{vid}^0, z_{aud}^0 means the first token of z_{vid}, z_{aud}, respectively, and FUSION$(\cdot)$ means a fusion model consisting of transformer encoder layers.

In the same manner, by forwarding $\alpha_{vid}(v_a^u), \alpha_{vid}(v_b^u)$ and a_a^u, a_b^u into the teacher model, we can calculate $\hat{y}_a$ and $\hat{y}_b$. Then, using the calculated $\hat{y}_a$ and $\hat{y}_b$ along with λ from Eq. 8, we compute the pseudo label $\bar{y}_{mix}$. This is used to optimize the model through the consistency loss as described in Eq. 12.

$$L_{mix} = \frac{1}{B_u}\sum^{B_u} \mathbb{1}(\max(\bar{y}_{mix}) \geq \tau)(\bar{y}_{mix} - \hat{y}_{mix})^2 \tag{12}$$

3.3 Audio Source Localization-guided Mixup

However, we observe that video and audio are augmented only within their respective modalities in the framework in Section 3.2. This approach does not consider the interrelation between the visual and audio information, even though they share the same video clip. Therefore, we propose the novel audio source localization-guided mixup. If the mask is generated through sampling from the audio source localization map and provided as the guide, the relationship between video and audio can be considered when performing the mixup.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

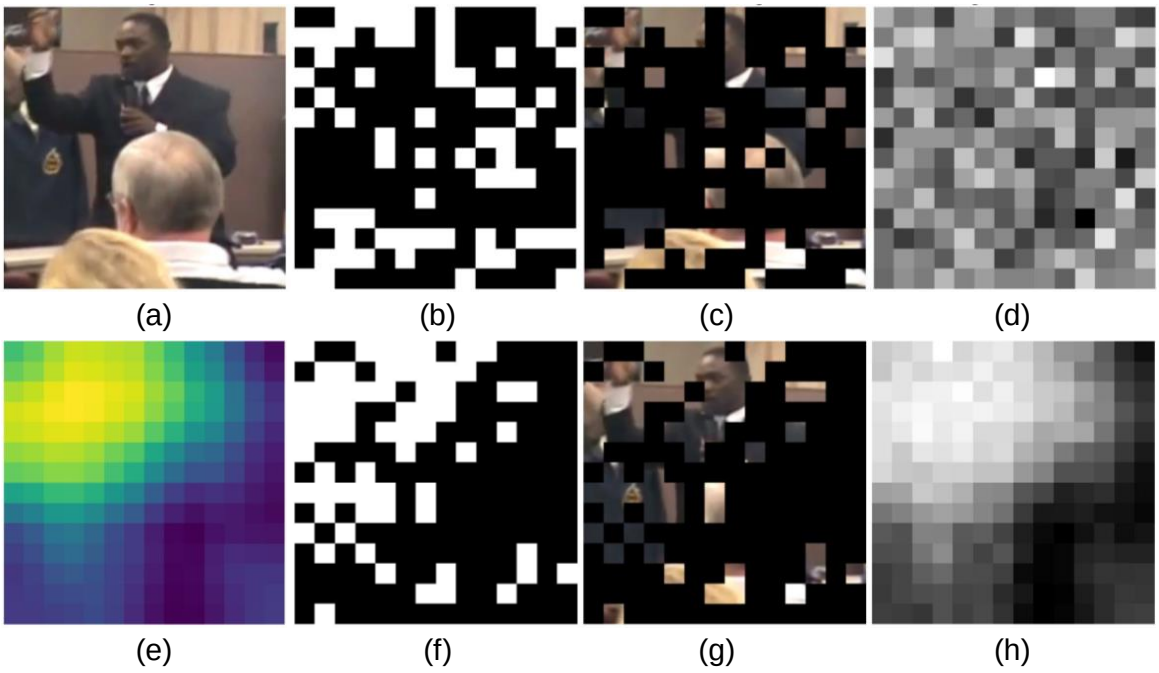


Figure 3: **Visualization of the TubeToken mask and our proposed Audio Source Localization-guided Mask.** When an original image (a) is provided, applying the TubeToken masking generates a mask with a random pattern (b). The created TubeToken mask is then utilized to produce a masked image, as shown in (c), which is used as input in SVFormer. Sampling this TubeToken mask 255 times allows for the observation of a random average TubeToken Mask, as seen in (d). In contrast, our proposed method involves creating the audio source localization-guided mask. This process starts with the original input (a), from which a localization map is generated, as shown in (e). This map is then employed as a weight for sampling, leading to the generation of the audio source localization-guided mask (f). Applying this mask to the original input results in a modified input, as depicted in (g). Furthermore, similar to the TubeToken mask, sampling the mask 255 times produces a result like (h). This process enables the capture of areas where the audio source is located, allowing us to consider the interrelation between the visual and audio modalities and perform the mixup accordingly.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

We introduce the audio source localization model that is composed of a visual encoder $E^{as}_{vid}(\cdot)$ and an audio encoder $E^{as}_{aud}(\cdot)$. For v^u and a^u sharing a video clip, $A_{vid}(v^u)$ is feed into $E^{as}_{vid}(\cdot)$, and $A_{aud}(a^u)$ is similarly feed into $E^{as}_{aud}(\cdot)$.

As a result, the visual feature f_{vid} and the audio feature f_{aud} are calculated. To compute their respective attention maps f^{attn}_{vid} and f^{attn}_{aud}, we perform matrix multiplication on the transposed features as shown in Eq. 13. Subsequently, we generate the final audio source localization map MAP using dot product operations as outlined in Eq. 14.

$$f^{attn}_{vid} = (f_{vid})(f_{vid})^T, \quad f^{attn}_{aud} = (f_{aud})(f_{aud})^T \tag{13}$$

$$\text{MAP} = (f^{attn}_{vid})(f^{attn}_{aud})^T \tag{14}$$

Building upon this, we can construct the novel mask M_{as} to replace the token-level mask M utilized in Eq. 7. We interpolate MAP to align with the dimensions of $Emb_{vid}(v^u)$, resulting in an interpolated map denoted as MAP$'$. By normalizing MAP$'$ with min-max normalization, we provide a basis for probabilistically selecting locations likely to contain the audio source. Employing multinomial distribution probabilities, we sample without replacement using MAP$'$ as weights. This process generates the new mask M_{as}, allowing us to retain $(\lambda \cdot N)$ tokens of $Embvid(v^u)$ for use. We visualize the differences between our proposed audio source localization-guided mask and the TubeToken mask in Figure 3.

3.4 Visual-Audio Contrastive Learning

To align visual and audio information, we employ a visual-audio contrastive learning loss. This approach utilizes the embeddings z^0_{vid} and z^0_{aud}, obtained from Eq. 9 and 10, respectively. Considering labeled, unlabeled, and mixed data, we define the video embeddings matrix $Z^l_{vid}, Z^u_{vid}, Z^m_{vid}$ in Eq. 15, where $i \in [B_l]$ and $j \in [B_u]$. A similar definition applies to the audio embeddings matrix Z_{aud}.

First, we perform L2 normalization on Z_{vid} and Z_{aud}. The similarity between video and audio is then computed using matrix multiplication as shown in Eq. 17, resulting in the visual-audio similarity matrix S.

$$Z_{vid} = [z^{l_{0,1}}_{vid}, \cdots, z^{l_{0,i}}_{vid} \ ; \ z^{u_{0,1}}_{vid}, \cdots, z^{u_{0,j}}_{vid} \ ; \ z^{m_{0,1}}_{vid}, \cdots, z^{m_{0,j}}_{vid}] \tag{15}$$

$$Z'_{vid} = \frac{Z_{vid}}{\|Z_{vid}\|_2}, \quad Z'_{aud} = \frac{Z_{aud}}{\|Z_{aud}\|_2} \tag{16}$$

$$S = \frac{1}{0.05} \cdot Z'_{aud}(Z'_{vid})^T \tag{17}$$

Subsequently, we compute the visual-audio contrastive loss L_c using the noise-contrastive estimation loss applied to S.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

$$\mathrm{NCE}_1 = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\frac{e^{S_{ii}}}{\sum_{j=1}^{K} e^{S_{ij}}}\right) \tag{18}$$

$$\mathrm{NCE}_2 = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\frac{e^{S_{ii}}}{\sum_{j=1}^{K} e^{S_{ji}}}\right) \tag{19}$$

$$L_c = \frac{\mathrm{NCE}_1 + \mathrm{NCE}_2}{2} \tag{20}$$

Here, K denotes the number of elements in Z_{vid} and Z_{aud}.

3.5 Training Objective

Our proposed model requires four main types of loss functions for training. These include the supervised learning loss (Eq. 1), the consistency regularization loss for unlabeled data (Eq. 6), and the contrastive loss for aligning visual and audio modalities (Eq. 20). Additionally, we utilize a consistency regularization loss for a mixture of unlabeled data, guided by audio source localization mixup (Eq. 12). The total loss used for training is formulated as follows:

$$L_{total} = L_s + \gamma_1 \cdot L_u + \gamma_2 \cdot L_{mix} + \gamma_3 \cdot L_c \tag{21}$$

where γ_1, γ_2, and γ_3 are the hyper-parameters for balancing the each loss.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

IV Experiment

In this section, we first explain the experimental setting. Next, we analyze the experimental results conducted with different numbers of labeled data on different datasets and discuss ablation studies.

4.1 Experimental Setting

Datasets We conduct experiments using three datasets: UCF-51 [48], Kinetics-400 [49], and VGGSound [50]. UCF-51 is a refined subset of the UCF-101 dataset, comprising 51 classes that include audio, consisting of 4.9K training samples and 1.4K test samples. Kinetics-400 covers 400 categories, each with approximately 10-second-long videos, amounting to 240K training samples and 20K validation samples. For this dataset, samples without audio are filtered out. VGGSound is composed of around 200K videos, each about 10 seconds in length, and provides annotations for 309 categories. After excluding unavailable videos, we have 183,730 training samples and 15,446 test samples. Before the experiments, the UCF-51, Kinetics-400, and VGGSound datasets are split to ensure each class contains 1 and 5 labeled data samples, respectively. Using 1 labeled data sample per class is a more extremely challenging experimental condition, which has also been used in semi-supervised image classification studies [15, 51]. Additionally, using 5 labeled data samples per class is similar to other studies [9, 11, 19] that take a 1% labeled samples setting, which uniformly uses only six labeled samples per class, despite the dataset being imbalanced.

Evaluation metric For evaluation, we use the accuracy of video action recognition.

Baselines We use the prior study SVFormer [9] and the state-of-the-art audio-visual semi-supervised action recognition, AvCLR [23], as our baselines. However, in the case of AvCLR, the implementation codes are not public so we report the reported results of their study. Additionally, we employ a model that modifies the fixed threshold used in the original SVFormer to a flexible threshold. The majority of the hyper-parameters are set up identically to those in SVFormer.

4.2 Implementation Details

For training, we follow the settings of SVFormer. All experiments are conducted with two NVIDIA A100 80GB GPUs. We utilize the FNAC [46] as the audio source localization model. FNAC is a two-stream network comprising two ResNet-18 models, serving as the visual and audio encoders, respectively. During the training phase, we use a version of FNAC that has been pre-trained on the Flickr-1K dataset [52] without object-guided localization. This model is kept frozen and is not further trained during our training phase. We set the ratio between labeled samples and unlabeled samples as 1:5 in the mini-batch following SVFormer's setting. Also, we use an SGD optimizer for training, with a momentum of 0.9 and a weight decay of 0.001. The values of γ_1 and γ_2 are set to 2, while for γ_3, we select the optimal value from $\{0.1, 0.2, 0.3\}$. Additionally, for the masking ratio λ, we sample from the beta

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

distribution Beta(α_1, α_2), setting α_1 to 5 and α_2 to 10. This is because guided by audio source localization, it is acceptable to take a smaller ratio of the video for creating the localization map due to the selection of significant regions. During the testing phase, the entire video is uniformly divided into five parts, and we perform cropping three times to cover most areas of the video clip, each with a size of 224×224. For the final testing prediction, we averaged a total of 15 predictions, obtained by performing the aforementioned three crops five times per video clip.

Model	Backbone	Input	Epoch	Threshold Type	Mask Type	UCF-51		Kinetics-400		VGGSound	
						51	255	400	2000	309	1545
AvCLR [23]	3D-Resnet50	V,A	800	Fix	-	-	50.1	-	-	-	-
SVFormer [9]	ViT-B	V	50	Fix	TubeToken	60.75	87.71	16.01	46.93	14.87	37.70
SVFormer [9]	ViT-B	V	50	Flex	TubeToken	63.58	86.63	16.69	47.62	15.98	36.72
Ours	ViT-B	V,A	50	Flex	ASL-Guided	**72.58**	**89.09**	**19.12**	**48.64**	**17.49**	**38.00**

Table 1: **Comparisons with state-of-the-art methods on UCF-51, Kinetics-400, and VGGSound.** Note that AvCLR is based on a CNN architecture, while SVFormer and our method are based on a Transformer architecture. Regarding inputs, 'V' represents the video modality, and 'A' denotes the audio modality. In threshold type, 'Fix' means fixed threshold proposed by FixMatch [12], and 'Flex' indicates flexible thresholding of FlexMatch [13]. In terms of mask type, 'TubeToken' refers to the TubeToken mask strategy proposed by SVFormer [9], and 'ASL-Guided' stands for Audio Source Localization-Guided, which is our proposed method. The numbers below each dataset indicate the total number of labeled data samples during the training phase.

4.3 Main Results

The main experimental results on UCF-51, Kinetics-400, and VGGSound datasets can be found in Table 1. Compared to previous methodologies, our proposed method demonstrates superior performance. Specifically, on the UCF-51 dataset, using only one labeled sample per class, we achieve an absolute performance improvement of 9.00% accuracy and also a relative accuracy improvement of 14.1% over the SVFormer, which only uses video modality with flexible thresholding. Moreover, on Kinetics-400, our model outperforms SVFormer by 2.43%p and 15.56%, and on VGGSound, we achieve 1.51%p and 9.45% higher performances in the absolute and relative accuracy improvement, respectively. Even though the gains become smaller when considering performance with the 5 samples per class which is easier, our approach still outperforms baselines and improves the performance. These results demonstrate the effectiveness of our proposed audio source localization-guided mixup method in considering the inter-modal relation between video and audio modalities.

Our method, which utilizes the visual-audio modality in a semi-supervised learning approach while also considering the inter-modal relation between video and audio, surpassed the state-of-the-art AvCLR, which is CNN-based, by 38.99%p on UCF-51. To demonstrate that the performance difference is not solely due to the backbone architecture, but also due to the effectiveness of our proposed methodology, additional experiments are explored in Section 4.4.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

Input	ASL Mask	Contrastive	Accuracy
V			63.58
V, A			66.72
V, A		✓	68.52
V, A	✓		71.55
V, A	✓	✓	**72.58**

Table 2: **Ablation study on the use of audio source localization-guided mask and contrastive learning.** 'ASL Mask' indicates the presence or absence of the audio source localization-guided mask (ASL Mask), while 'contrastive' refers to the use or non-use of visual-audio contrastive learning. This experiment is conducted on the UCF-51 dataset using one labeled sample per class. It aims to understand the effectiveness of our proposed ASL mask and audio-visual contrastive learning.

4.4 Ablation Studies

To understand the impacts of the methods we propose, we conduct the ablation studies on UCF-51.

Analysis of Mask Type and Contrastive Learning We conduct an ablation study to evaluate the effectiveness of our proposed audio source localization-guided mask (ASL mask) and visual-audio contrastive learning. When the ASL mask is not used, the TubeToken mask is employed. First, the results show that using both video and audio modalities leads to a performance improvement of 3.14 %p. Furthermore, utilizing visual-audio contrastive learning results in an additional performance increase of 1.80%p. Notably, using our proposed ASL mask results in a performance improvement of 4.83%p over the TubeToken mask. When combined with contrastive learning, this method achieves an accuracy of 72.58%. These findings demonstrate the effectiveness of our audio source localization-guided mixup in considering the inter-modal relation between video and audio modalities, while the visual-audio contrastive learning contributes to further performance enhancements.

Threshold for Pseudo Label We conduct an ablation study to assess the impact of the threshold τ for generating pseudo labels in our proposed audio source localization guided-mixup method. In this study, the predicted probability calculated by the teacher model is used as a pseudo label, and we explore how different τ values affect this process.

During the training phase on the UCF-51 dataset, only one labeled sample per class is utilized. We observed that a lower threshold of 0.1 generally results in most predicted probabilities being used as pseudo labels. However, this can lead to performance degradation due to the accumulation of bias from low-confidence pseudo-labels.

Moreover, ours of the flexible thresholding method proposed in FlexMatch [13] allows us to maintain superior performance compared to SVFormer, which also incorporates audio modality, even as τ increases. Among various τ values, we identified 0.3 as the optimal value and set it for our main experi-

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

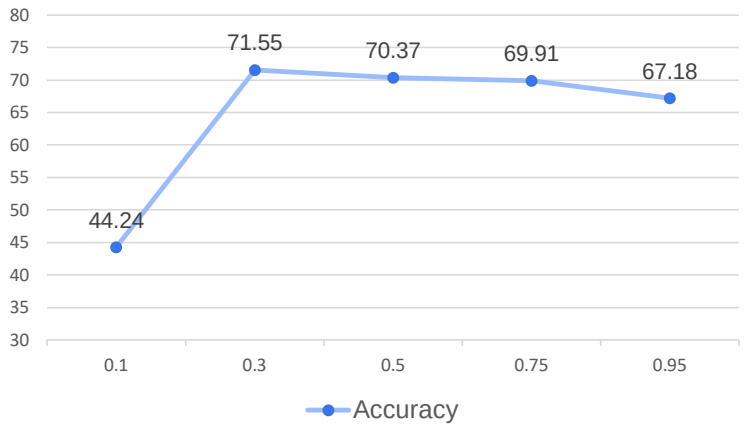


Figure 4: **Performance Impact of Hyper-parameter τ.** This figure presents the results of experiments conducted to explore the performance impact of the hyper-parameter τ. It specifically focuses on the changes in performance with varying thresholds of τ during training with only one labeled sample in the UCF-51 dataset.

ments and other ablation studies. The performance variations according to different τ values can be seen in Figure 4.

Mask Sampling per Varying Frames Inspired by the applications of tube-shaped masks in the video domain [1, 9], which utilize consistent tokens across time to prevent information leakage from adjacent frames, we conduct an ablation study on the creation of audio source localization masks with the one labeled sample per class on UCF-51. This study involves averaging maps generated from adjacent frames and using these averages for sampling to create masks. Our experiments are designed to test averages over 1, 2, 4, 8 frames. The results of this experiment can be seen in Figure 5.

From our findings, we observe that contrary to the TubeToken mask approach, generating masks based on each individual frame achieves the best performance for our proposed audio source localization-guided masks. This can be attributed to the fact that, while audio source localization maps are generated for each frame based on both visual and audio information, the visual information varies from frame to frame. Therefore, when using the average of audio source localization maps generated from different frames, it can prevent information leakage from adjacent frames during token-level mixup, similar to the TubeToken approach. This method might be effective for static videos where the audio source location changes minimally. However, for videos with rapid changes, this averaging approach uses a mean of different audio source locations, which may not accurately represent the true location in dynamic scenes. Consequently, this can lead to significant information loss and, ultimately, a decrease in performance.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

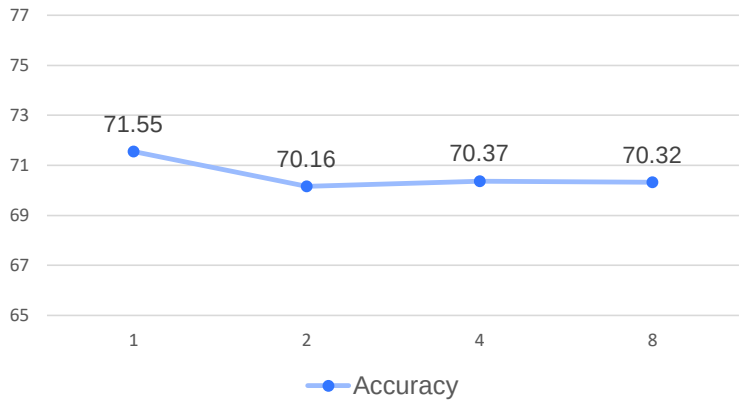


Figure 5: **Ablation study on the effect of varying frame counts (1, 2, 4, and 8 frames) on the audio source localization map.** This study evaluates the impact of averaging the localization maps over different numbers of frames before using them in the sampling process.

V Conclusion

In this paper, we introduce a transformer-based semi-supervised multimodal video action recognition approach. We expand the previous visual framework to the visual-audio framework to leverage the visual-audio information available from video clips. Furthermore, to overcome the limitations of existing augmentation methods that consider individual modalities, we propose the audio source localization-guided mixup method. This approach considers the interrelation between visual and audio information. Our proposed method demonstrates superior performance over the existing state-of-the-art methods on the UCF-51, Kinetics-400, and VGGSound datasets. As a limitation, our approach depends on the audio source localization method, but we expect the performance can be improved as the following audio source localization is improved.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

References

[1] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.

[2] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[3] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao, "Uniformerv2: Unlocking the potential of image vits for video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1632–1643.

[4] Y. Liu, K. Wang, L. Liu, H. Lan, and L. Lin, "Tcgl: Temporal contrastive graph for self-supervised video representation learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 1978–1993, 2022.

[5] T. Yang, Y. Zhu, Y. Xie, A. Zhang, C. Chen, and M. Li, "Aim: Adapting image models for efficient video action recognition," *arXiv preprint arXiv:2302.03024*, 2023.

[6] J. Chen and C. M. Ho, "Mm-vit: Multi-modal video transformer for compressed video action recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 1910–1921.

[7] S. N. Gowda, M. Rohrbach, F. Keller, and L. Sevilla-Lara, "Learn2augment: learning to composite videos for data augmentation in action recognition," in *European conference on computer vision*. Springer, 2022, pp. 242–259.

[8] A. Tong, C. Tang, and W. Wang, "Semi-supervised action recognition from temporal augmentation using curriculum learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1305–1319, 2022.

[9] Z. Xing, Q. Dai, H. Hu, J. Chen, Z. Wu, and Y.-G. Jiang, "Svformer: Semi-supervised video transformer for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 816–18 826.

[10] J. Xiao, L. Jing, L. Zhang, J. He, Q. She, Z. Zhou, A. Yuille, and Y. Li, "Learning from temporal gradient for semi-supervised action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3252–3262.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

[11] Y. Xu, F. Wei, X. Sun, C. Yang, Y. Shen, B. Dai, B. Zhou, and S. Lin, "Cross-model pseudo-labeling for semi-supervised action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2959–2968.

[12] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[13] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 408–18 419, 2021.

[14] B. Chen, J. Jiang, X. Wang, P. Wan, J. Wang, and M. Long, "Debiased self-training for semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 424–32 437, 2022.

[15] Y. Wang, H. Chen, Q. Heng, W. Hou, Y. Fan, Z. Wu, J. Wang, M. Savvides, T. Shinozaki, B. Raj *et al.*, "Freematch: Self-adaptive thresholding for semi-supervised learning," *arXiv preprint arXiv:2205.07246*, 2022.

[16] Y. Chen, X. Tan, B. Zhao, Z. Chen, R. Song, J. Liang, and X. Lu, "Boosting semi-supervised learning by exploiting all unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7548–7557.

[17] I. R. Dave, M. N. Rizve, C. Chen, and M. Shah, "Timebalance: Temporally-invariant and temporally-distinctive video representations for semi-supervised action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2341–2352.

[18] L. Jing, T. Parag, Z. Wu, Y. Tian, and H. Wang, "Videossl: Semi-supervised learning for video classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1110–1119.

[19] B. Xiong, H. Fan, K. Grauman, and C. Feichtenhofer, "Multiview pseudo-labeling for semi-supervised learning from video," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7209–7219.

[20] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[21] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

[22] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[23] M. Assefa, W. Jiang, J. Zhan, K. Gedamu, G. Yilma, M. Ayalew, and D. Adhikari, "Audio-visual contrastive and consistency learning for semi-supervised action recognition," *IEEE Transactions on Multimedia*, 2023.

[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[30] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[31] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.

[32] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[33] W. Dong-DongChen and Z. WeiGao, "Tri-net for semi-supervised deep learning," in *Proceedings of twenty-seventh international joint conference on artificial intelligence*, 2018, pp. 2014–2020.

[34] F. Mao, X. Wu, H. Xue, and R. Zhang, "Hierarchical video frame sequence representation with deep convolutional graph network," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

[35] S. Bhardwaj, M. Srinivasan, and M. M. Khapra, "Efficient video classification using fewer frames," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 354–363.

[36] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4804–4814.

[37] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[38] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[39] M. Assefa, W. Jiang, K. Gedamu, G. Yilma, D. Adhikari, M. Ayalew, A. Mohammed, and A. Erbad, "Actor-aware self-supervised learning for semi-supervised video representation learning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[40] J. Wu, W. Sun, T. Gan, N. Ding, F. Jiang, J. Shen, and L. Nie, "Neighbor-guided consistent and contrastive learning for semi-supervised action recognition," *IEEE Transactions on Image Processing*, 2023.

[41] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, "Learning to localize sound source in visual scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4358–4366.

[42] A. Senocak, H. Ryu, J. Kim, T.-H. Oh, H. Pfister, and J. S. Chung, "Sound source localization is all about cross-modal alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7777–7787.

[43] S. J. Um, D. Kim, and J. U. Kim, "Audio-visual spatial integration and recursive attention for robust sound source localization," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3507–3516.

[44] X. Hu, Z. Chen, and A. Owens, "Mix and localize: Localizing sound sources in mixtures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 483–10 492.

[45] S. Mo and P. Morgado, "Localizing visual sounds the easy way," in *European Conference on Computer Vision*. Springer, 2022, pp. 218–234.

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

[46] W. Sun, J. Zhang, J. Wang, Z. Liu, Y. Zhong, T. Feng, Y. Guo, Y. Zhang, and N. Barnes, "Learning audio-visual source localization via false negative aware contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6420–6429.

[47] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[48] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[49] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[50] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[51] T. Lucas, P. Weinzaepfel, and G. Rogez, "Barely-supervised learning: Semi-supervised learning with very few labeled images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1881–1889.

[52] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY