Enhancing Audio-Visual Association with Self-Supervised Curriculum Learning

Jingran Zhang[1], **Xing Xu**[1*], **Fumin Shen**[1], **Huimin Lu**[2], **Xin Liu**[3], **Heng Tao Shen**[1]
[1]Center for Future Multimedia and School of Computer Science and Engineering, University of Electronic Science and Technology of China, China
[2]Kyushu Institute of Technology, Japan
[3]Huaqiao University, China
{jrzhang339, fumin.shen}@gmail.com, xing.xu@uestc.edu.cn, shenhengtao@hotmail.com, dr.huimin.lu@ieee.org, xliu@hqu.edu.cn

Abstract

The recent success of audio-visual representations learning can be largely attributed to their pervasive concurrency property, which can be used as a self-supervision signal and extract correlation information. While most recent works focus on capturing the shared associations between the audio and visual modalities, they rarely consider multiple audio and video pairs at once and pay little attention to exploiting the valuable information within each modality. To tackle this problem, we propose a novel audio-visual representation learning method dubbed *self-supervised curriculum learning* (SSCL) under the teacher-student learning manner. Specifically, taking advantage of contrastive learning, a two-stage scheme is exploited, which transfers the cross-modal information between teacher and student model as a phased process. The proposed SSCL approach regards the pervasive property of audiovisual concurrency as latent supervision and mutually distills the structure knowledge of visual to audio data. Notably, the SSCL method can learn discriminative audio and visual representations for various downstream applications. Extensive experiments conducted on both action video recognition and audio sound recognition tasks show the remarkably improved performance of the SSCL method compared with the state-of-the-art self-supervised audio-visual representation learning methods.

Introduction

The co-occurrence of acoustic signal and visual appearance provides potential cues for humans experiencing the world. For example, while hearing ball bouncing, we can match it to the scenario of basketball games from numerous visual scene candidates. The concurrency is an inherent property that sound is a kind of vibration generated by surrounding objects (Hu, Nie, and Li 2019), and exists through our daily life, such as the crowd cheering and laughing, announcer speaking, whistling, and ball bouncing, *etc*. For machine models, these inherent and pervasive correspondences raise the possibility to possess similar abilities like humans by investigating audio-visual representation learning and discovering their complex correlations with different audio-visual messages. Additionally, in contrast to expensive human-annotation, the concurrent audiovisual message provides free and pervasive supervised signal for exploring self-supervision learning and co-training a multi-modality network on large-scale unlabeled data.

The recent studies on audio-visual representations learning can be generally categorized into two types: *Audio-Visual Correspondence (AVC)* (Arandjelovic and Zisserman 2017) and *Audio-Visual Synchronization (AVS)* (Chung, Chung, and Kang 2019). The two types are mainly about setting up a verification task that predicts whether an input pair of an audio and an video clip is matched or not. The positive audio and video pairs are typically sampled from the same video. The main difference between AVC and AVS is how to treat the negative audio and video pair, *i.e.,* the negative pair in AVC is mostly constructed by audio and video from different videos while in AVS is to detect the misalignments between negative audio and video pair from the same video. To escape from the relying on stronger supervision signal, existing studies attempt to address those problems on the aspect of cross-modal knowledge transfer (Aytar, Vondrick, and Torralba 2016; Owens et al. 2018) to directly predict the correspondence of the audio-visual messages. Typically, a two-stream audio-visual model is trained in (Arandjelovic and Zisserman 2017; Korbar, Tran, and Torresani 2018a) to judge the pair just with the given audio-visual correspondence. Nevertheless, those works mainly consider the information shared between two modalities for semantic representations learning, but neglect the important cues of multiple audio and video pairs at once. Besides, they also rarely consider exploiting the useful information underlying the same modality to model the data distribution.

To address the above issues, we consider to learn the correspondence between audio and visual from the pipeline of teacher-student learning. Specifically, a teacher network will teach the student network to obtain semantic audio and visual representations in unlabeled video with the contrastive learning manner. Previous works (Aytar, Vondrick, and Torralba 2016; Arandjelovic and Zisserman 2017) on teacher and student knowledge transfer mainly focus on mimicking the intermediate representations or logits of teacher networks in a pairwise manner. However, the concurrency knowledge of audio and video derived from the same video

*Corresponding author

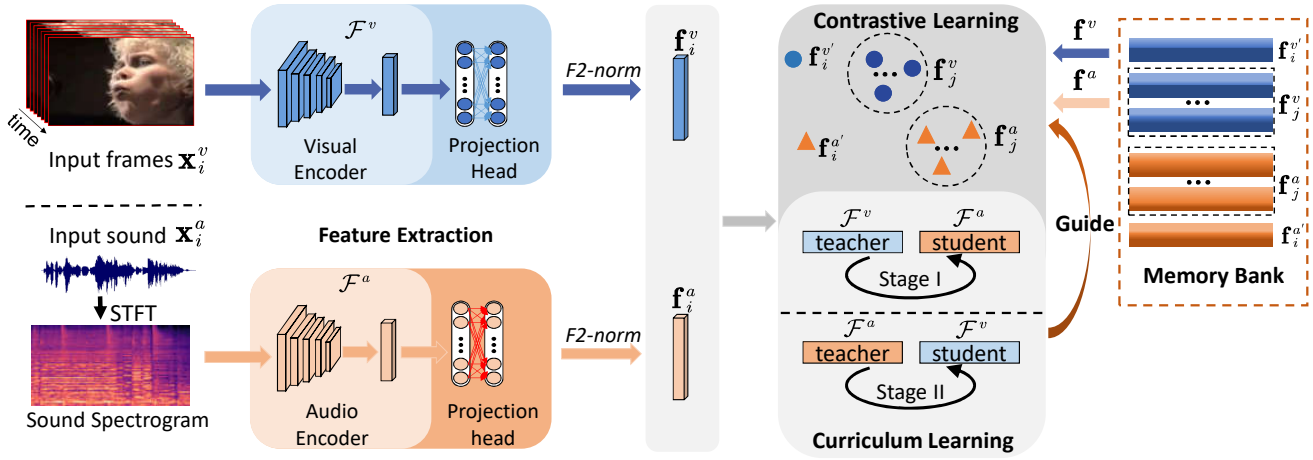

Figure 1: Illustration of the General Framework of Our Proposed SSCL Method for Audio-Visual Representation Learning.

may only reflect a single aspect of the complete knowledge encapsulated in a cumbersome network. It is intuitive that the intermediate representations of the teacher network could provide more discriminative knowledge. To this end, we can naturally exploit the contrastive learning to capture richer structured knowledge from the self-supervision prediction of the teacher model. This scheme not only helps the student model to obtain the knowledge of how audio and visual objects are concurrent but also reveals why the unmatched noisy audio and video pairs are different by contrast to the matched pair.

A great challenge is that, in a heterogeneous complexity of audio-visual scenes, directly transferring the information from one network to another may deteriorate the contrastive learning process. For example, in a basketball game or a football game, one cannot distinguish it from the crowd cheering and announcer speaking without hearing ball bouncing or kicking. Therefore, in this paper, we propose a novel Self-Supervised Curriculum Learning (SSCL) method to distinguish uni-modality instance and transfer cross-modal correlation information to each others. As the general framework of SSCL illustrated in Figure 1, the curriculum learning process has two successive stages. The visual model is exploited as a teacher model in stage-I with contrastive learning to enhance the feature learning procedure, and then the cross-modal training is deployed to transfer knowledge to the student model. While in the stage-II, the roles of the teacher and the student models that are exploited in the stage-I are exchanged for conducting the cross-model transfer process. More specifically, the visual encoder (3D ConvNets) and audio encoder (2D ConvNets) are first used to process the pairwise frames and sound input $(\mathbf{x}_i^v, \mathbf{x}_i^a)$ into a respective feature embedding $(\mathbf{f}_i^v, \mathbf{f}_i^a)$. Note that the sound is transformed into a spectrogram image. Afterward, the feature embedding pair is projected into two separate 128-D embedding spaces. Finally, the contrastive learning based audio and visual knowledge transfer process is guided by our proposed two-stage curriculum learning scheme. Meanwhile, the memory bank which achieves a moving average of representation is exploited to store negative samples for contrastive learning. Notably, the whole framework takes the concurrency of the acoustic signal and visual appearance as the supervision for training. To this end, we evaluate the audio-visual encoder on various audio-visual downstream tasks, *e.g.,* action recognition and audio recognition, In experiment section , we conduct extensive experiments on those tasks and demonstrate the superiority of our SSCL model compared to a bundle of state-of-the-art self-supervised learning methods for audio-visual representation learning.

The main contributions of this paper are summarized as:

- We propose a self-supervised audio-visual modality transfer framework termed SSCL to explore more coherent knowledge from a teacher network to a student network, where contrastive learning is leveraged to capture the correspondence between audio and visual information.
- We develop a two-stage curriculum learning process to reason about multiple single-modality instances and distill cross-modal correction information. This process not only improves the overall distillation performance but also regularizes the teacher and student model to generalize on noisy and complex scenarios.
- We further apply the learned audio-visual representations to a variety of audio and visual downstream tasks. The extensive experiments verify the powerful audio-visual representations learned by our SSCL method, leading to the remarkable improvement of the performance on the downstream tasks compared with previous approaches.

Related Work

Self-Supervised Representation Learning of Audio-Visual Data There has been increased interest in learning the audio-visual representation from a video to improve the performance of audio and visual models or solve data shortage problems (Arandjelovic and Zisserman 2017; Korbar, Tran, and Torresani 2018b; Sayed, Brattoli, and Ommer

2018). Self-supervised learning, which does not require human annotation, has been leveraged to train both audio and visual networks for multimodal audio-visual representation learning. Generally, self-supervised learning has the potentiality to leverage a large amount of unlabeled data because it proposes a pretext task to generate pseudo labels and to explore data structure. A wide range of pretext tasks have been proposed, like colorization (Zhang, Isola, and Efros 2016), rotation prediction (Gidaris, Singh, and Komodakis 2018), *etc*, which are usually based on the form of contrastive loss functions (Tian, Krishnan, and Isola 2019).

For audio-visual representation learning, most recent works (Korbar, Tran, and Torresani 2018b; Owens and Efros 2018; Alwassel et al. 2019; Gan et al. 2020) exploited the co-occurrence of audio wave and visual object to learn more compact representations, which are beneficial for many downstream applications, *e.g.,* sound classification (Piczak 2015a; Gao et al. 2020), separation (Gao, Feris, and Grauman 2018; Zhao et al. 2018) and localization (Kidron, Schechner, and Elad 2005), visual representation (Hu, Nie, and Li 2019; Gan et al. 2019; Shukla, Petridis, and Pantic 2020), and synchronization (Korbar, Tran, and Torresani 2018a). In particular, the audio or visual information has shown to be useful as supervision for pre-training visual/audio models. For example, in (Owens and Efros 2018), an early-fusion multi-sensory network is trained to predict whether video frames and audio are temporarily aligned, and in (Khosravan, Ardeshir, and Puri 2019), an attention mechanism together with a two-stream network structure is developed to localize the sound source.

Cross-Modal Learning and Distillation. A large range of modalities from audio, visual and optical flow, are inherent in videos, and their correspondence can be used as pseudo supervisory signal for representation learning. Multimodal learning aims to leverage correction from the multiple signals of the same source and has increased widely interest in supervised and unsupervised learning scenarios (Wang et al. 2017; Shen et al. 2020; Xu et al. 2020b). The typical pipelines of the common approaches include: training the multi-modal in the same procedures (Pham et al. 2019; Xu et al. 2020), maximizing the mutual information across those modalities of unlabeled data using a semi-supervised method (Tian, Krishnan, and Isola 2019), designing different modules for different modalities with a separate learning process, and common latent subspace learning for all modalities of the data. Furthermore, model-transfer based methods are also widely studied for knowledge transfer between different modalities (Aytar, Vondrick, and Torralba 2016; Owens et al. 2018).

The core idea of knowledge transfer is to transfer discriminative knowledge from a well-trained complex model to a simple model. Typically, the teacher and student model were exploited in this literature. Several works (Aytar, Vondrick, and Torralba 2016; Gupta, Hoffman, and Malik 2016) trained an encoder on one modality and transferred its discriminative knowledge to the encoder from another modality in a supervised manner. However, those methods generally rely on label-dependent features so that the teacher trained with such supervisory signals can be directly transferred to the target student model. Therefore, further works attempt to investigate the specific property from unlabeled data to obtain the benefit from the structure of data. Recently, several works have proposed to leverage the natural correspondence (Owens and Efros 2018) and synchronization (Korbar, Tran, and Torresani 2018b) between the audio and visual modality for effective multi-modal representations learning. Other methods (Alwassel et al. 2019) directly modeled the audio-visual correspondence without teacher supervision by predicting whether the sounds and frames are from the same video or not.

Our proposed SSCL is also inspired by the existing studies on cross-modal learning. However, those scene-level audio-visual consistencies typically lack specific annotations and suffer from the defects of inefficiency and irrelevance in the noisy scene, *e.g.,* yackety-yak, noise source, and insignificant visual background in any scenarios. Our work is totally self-supervised, which trains on a larger amount of unlabeled video and transfers well to a wide range of downstream tasks. Additionally, even faced with complex audio-visual scenarios, our method can extract the audio-visual representation with the help of the synchronization between genuine vision and sound.

Methodology

Overview of Our SSCL Approach

Suppose we have a video dataset (N sample) $\mathcal{V} = \{\mathbf{V}_i\}_{i=1}^{N}$, and a visual encoder $\mathcal{F}^v$ and audio encoder $\mathcal{F}^a$, an unlabelled video clip $\mathbf{V}_i = \{\mathbf{v}_1^i, \mathbf{v}_2^i, \ldots, \mathbf{v}_T^i\}$ is processed as a pair representations $\mathbf{f}_i = (\mathbf{f}_i^v, \mathbf{f}_i^a)$, where the T is the clip length, and the $\mathbf{f}_i^v$ is the visual representation extracted by the visual encoder $\mathcal{F}^v$ and the $\mathbf{f}_i^a$ is the audio representation extracted by the audio encoder $\mathcal{F}^a$. Our goal is to effectively train a visual and audio encoder $\mathcal{F}^v, \mathcal{F}^a$ and make it possess the ability to generate uni-modal representation $\mathbf{f}_i^v, \mathbf{f}_j^a$ and obtain effective cross-modal perception $(\mathbf{f}_i^v, \mathbf{f}_i^a)$ by exploiting the correlation of audio and visual within each video clip. The resulting representations are close for similar videos while distinguishable for dissimilar videos.

To accomplish this goal, our proposed SSCL approach aims to transfer information from the audio representation $\mathbf{f}^a$ to the visual representation $\mathbf{f}^v$, and vice versa. As illustrated in Figure 1, a teacher network $\mathcal{F}_\mathcal{T}$ is trained on a source modality $\mathcal{D}_\mathcal{T}$, and then transferred the knowledge to a student network $\mathcal{F}_\mathcal{S}$ and adapted it to another modality $\mathcal{D}_\mathcal{S}$. Note that the features of the teacher network are still valuable to help with the learning of the student in another domain. In a typical transfer task (Hinton, Vinyals, and Dean 2015), the loss is defined as:

$$\mathcal{L} = -\sum_{\mathbf{V}_i \sim \mathcal{V}} \sum_{c=1}^{C} \mathbb{P}_c^t(\mathbf{V}_i; \tau) \log(\mathbb{P}_c^s(\mathbf{V}_i; \tau)), \qquad (1)$$

where $\mathbb{P}_k^t$ and $\mathbb{P}_k^s$ denote the output probability of teacher and student network at class c, C is the total number of classes and τ is the temperature parameter. Specifically, it encourages the student network output to be as similar as possible to the teachers' by minimizing the Kullback-Leibler (KL) divergence between their outputs.

However, such a transfer is conducted on an unlabelled video dataset $\mathcal{V}$ without any ground-truth labels for the original training task on the source modality. Fortunately, the dataset $\mathcal{V}$ can be processed as a paired dataset $\mathcal{V} = \{\mathbf{V}_i = (\mathbf{x}_i^v, \mathbf{x}_i^a) | \mathbf{x}^v \in \mathcal{D}^v, \mathbf{x}^a \in \mathcal{D}^a\}_{i=1}^N$. Inspired by the works (Tian, Krishnan, and Isola 2020; Xu et al. 2020a), we exploit the contrastive loss (Hénaff et al. 2019; Tian, Krishnan, and Isola 2019; Chen et al. 2020) to match the features of the source domain $\mathcal{D}_\mathcal{T}$ and targeted domain $\mathcal{D}_\mathcal{S}$:

$$\mathcal{L}_{\text{ctr}}(\mathbf{f}_i^s, \mathbf{f}^t) = -\sum_{\mathbf{f}_k^t \sim \mathcal{D}_\mathcal{T}} \mathbb{E}\left[\log \frac{\exp(\text{sim}(\mathbf{f}_i^s, \mathbf{f}_i^t)/\tau)}{\sum_{j=1}^{K+1} \exp(\text{sim}(\mathbf{f}_i^s, \mathbf{f}_j^t)/\tau)}\right], \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ is a function that measures the similarity of the two input terms (dot product is adopted here), K is the number of negative samples and τ is a temperature parameter. This loss can maximize the similarity between positive pairs $(\mathbf{f}_i^s, \mathbf{f}_i^t)$ and minimize the similarity between negative pairs $(\mathbf{f}_i^s, \mathbf{f}_j^t), j \neq i$. Moreover, minimizing the objective loss in Eq. 2 encourages high mutual information between student representation $\mathbf{f}^s$ and teacher representation $\mathbf{f}^t$.

Curriculum Learning

Generally, it is hard to directly optimize Eq. 2 by training from scratch, because the audio and visual information might be noisy and irrelevant. We found that if we train both models simultaneously, the noisy information will destroy the updating process at the beginning. The latter experiment proves that the test results are consequently poor while optimizing the objective from scratch. Therefore, we need to design an effective training strategy to transfer the semantic correction information and suppress the noisy information between the two networks. To this end, we present a curriculum learning strategy (Korbar, Tran, and Torresani 2018b; Hu et al. 2020) by pre-training the teacher model.

Our curriculum learning strategy consists of two stages. Specifically, in the stage-I, we fix the student model and only update the parameters of the teacher model with a *self-instance discriminator*, and then jointly train the teacher-student models according to Eq. 2. While in the stage-II, we exchange the role of teacher and student model and fix the original teacher model and update the parameters of original student model in the same way, and after that, we jointly train the teacher-student model again. Notably, we can perform the two stages recursively.

The stage-I is to pre-train the teacher network directly at the beginning. Similar to the cross-modal contrastive loss defined in Eq. 2, we present the pre-training of teacher model loss as follows:

$$\mathcal{L}_{\text{ctr}}(\mathbf{f}_i^t, \mathbf{f}^t) = -\sum_{\mathbf{f}_k^t \sim \mathcal{D}_\mathcal{T}} \mathbb{E}\left[\log \frac{\exp(\text{sim}(\mathbf{f}_i^t, \mathbf{f}_i^{t'})/\tau)}{\sum_{j=1}^{K+1} \exp(\text{sim}(\mathbf{f}_i^t, \mathbf{f}_j^t)/\tau)}\right], \tag{3}$$

Where we only take uni-modal pair, $\mathbf{f}_i^{t'}$ is the feature of transformed sample $\mathbf{x}_i^t$. This pre-training process is also seen as a *self-instance discriminator* (Wu et al. 2018) by directly optimizing in the teacher's feature space. Fine-tuning the student model after the teacher model is well initialized will further boost the jointly updating process.

In our self-instance discriminator, each positive pair $(\mathbf{f}_i^t, \mathbf{f}_i^{t'})$ has K negative pairs $(\mathbf{f}_i^t, \mathbf{f}_j^t)$. The positive sample is obtained with the model by applying a set of transformers to sample $\mathbf{f}_i^t$. While the negative samples $\mathbf{f}_j^t$ are uniformly drawn from the dataset $\mathcal{D}_\mathcal{T}$ excluding sample $\mathbf{f}_i^t$. For visual transformation, except the traditional image spatial transformations like center cropping, randomly horizontal flipping, randomly color jitter, and randomly gray-scale, we also adopt temporal jitter that treats nearby clips from the same video as counterparts. The whole transformation process can be expressed as follows:

$$\mathbf{f}^{t'} = \mathcal{F}_\mathcal{T}(\text{Tmp}(\sum_j \text{Spa}(\sum_{i=1+j}^{T+j} \mathbf{x}_i^t)), \tag{4}$$

where the $\text{Spa}(\cdot)$ represents the set of image spatial transformations that are applied on each frame, and the $\text{Tmp}(\cdot)$ represents the temporal jitter, j is the temporal sliding step and randomly generated. For audio, we exploit the augmentation method proposed in this work (Park et al. 2019), which consists of *wrapping the features*, *masking blocks of frequency channels*, and *masking blocks of time steps*.

Using Memory Bank

From the previous section, the contrastive loss is the main objective function to transfer semantic related audio-visual information and capture audio or visual instance-level discrimination information. Prior works (van den Oord, Li, and Vinyals 2018; Hénaff et al. 2019) have proven that a larger number of negative samples can increase the contrastive learning performance. However, it is with high consumption to obtain a large number of negatives when increasing the batch to a large size. To reduce for computation capacity with larger batch size while using the standard optimizer of stochastic gradient descent (SGD), the memory-bank technique is widely used in this literature for caching features (Wu et al. 2018; He et al. 2020).

Memory Bank. In the pre-training process of each stage of curriculum learning, we use the *dictionary queue* $\mathcal{Q}$ proposed in (He et al. 2020) as the bank to cover negative features $\{\mathbf{f}_j^t\}_{j=1,j\neq i}^{K+1}$ for each input feature $\mathbf{f}_i^t$. The queue is dynamically evolving during training with a *momentum contrast* they proposed and in some sense represents a sampled subset of all data $\mathcal{D}_\mathcal{T}$.

In the joint-training process of each stage of curriculum learning, to transfer the knowledge from the teacher network to the student, we use a memory bank $\mathcal{M}$ to store all the feature representations $\{\mathbf{f}_j^t\}_{j=1}^N$ to match feature $\mathbf{f}_i^s$. The representations are momentum updated with prior epoch representations. Note that both ways allow us to replace negative samples with the bank representations without increasing the training batch size.

Final Loss. The Noise-Contrastive Estimation (Wu et al. 2018; Tian, Krishnan, and Isola 2019) and simple softmax-based classifiers (He et al. 2020) are widely used to approximate the full softmax distribution. We directly apply the

$K+1$-way softmax-based classifier to estimate Eq. 3 according to the dictionary queue $\mathcal{Q}$. Although we cache all the features in the memory bank $\mathcal{M}$, directly computing Eq. 2 is prohibitive as the dataset size N is very large. To tackle the computation issue of the similarity measure among all the instances in the set, we can efficiently retrieve K negative samples from the memory bank $\mathcal{M}$ to match with each positive pair $(\mathbf{f}_i^s, \mathbf{f}_i^t)$. We also adopt a simple $K+1$-way softmax-based classification loss as follows:

$$\mathcal{L}_{\text{sof}}(\mathbf{f}_i^s, \mathbf{f}^t) = -\mathbf{f}_i^{s^T} \cdot \mathbf{f}_i^t/\tau + \log(\sum_{j=1}^{K+1} \exp(\mathbf{f}_i^{s^T} \cdot \mathbf{f}_j^t/\tau)), \quad (5)$$

where we only show the transfer type. The objective in self-discriminator just replaces the $\mathbf{f}^s$ term. This loss can preserve information of matched pairs by optimizing the contrastive learning objectives.

Experiments

Experimental Setup

We follow the common practice in self-supervised learning (Xu et al. 2019; Alwassel et al. 2019) and evaluate the performance of SSCL in the downstream transfer-learning experiments. We conduct experiments on a variety of tasks, focusing on *action recognition* for visual representation evaluation and *sound recognition* for audio representation evaluation. The empirical experiments cover: 1) a fine-tuned setting during transfer in which the parameters of the encoder obtained with self-supervised training are employed for action recognition, thus evaluating the encoder "initialization" and 2) a linear probe setting during transfer in which the parameters of the encoder except the extra linear classification layer are fixed, thus using the encoder as a feature extractor for sound classification.

Pre-training Dataset For audio-visual pre-training, the standard dataset, **Kinetics-400** (Kay et al. 2017), is exploited as an unlabeled benchmark to pre-train our model. The Kinetics-400 dataset is also a widely used dataset for self-supervised audio-visual representation learning. It consists of 306,000 video clips available on YouTube website and covers 400 human action classes, including human-object interactions as well as human-object interactions. The average duration of a video clip is around 10 seconds and there are at least 600 video clips per action class.

Video and Audio Encoder We apply the S3D (Xie et al. 2018) network as the video encoder and a 10-layers ResNet (He et al. 2016) as the Audio encoder to extract visual and audio features respectively. Note that visual and audio features of each encoder are projected into two fully connected layers with an intermediate size of 512-D to produce 128-D embeddings as in which are normalized by a L2 Normalization. The 128-D embeddings are used for contrastive loss. This is fair with the related work (Sun et al. 2019).

Training Details To extract visual features, we sample 16 frames of a video clip at a sliding step 4 (around 3 seconds) and resize the frame to 112×112 resolutions. While for audio features extraction, we randomly sample 2 seconds of audio and compute a log spectrogram of size 128×128 (128 times steps with 128 frequency bands). The model is training with SGD using a linear warm-up scheme at an initial learning rate of 0.03. The SGD weight decay is 10^{-5} and the momentum is 0.9. The total epochs we used are 200 and the batch size is set as 128 with experiments on 8 GPU cards. We set the negative pairs K as 16,384, the temperature parameter τ as 0.07.

Downstream Tasks In this section, to investigate the correlation between self-supervised audio-visual learning and downstream tasks, we evaluate the quality of the pre-training audio-visual representations by transferring it to action and sound recognition. We evaluate the *visual representation* $\mathbf{f}^v$ with action recognition on the UCF-101 (Soomro, Zamir, and Shah 2012) and the HMDB-51 (Kuehne et al. 2011) datasets. Moreover, we also evaluate the *audio representation* $\mathbf{f}^a$ with sound classification on the ESC-50 (Piczak 2015b) and the DCASE (Stowell et al. 2015) datasets.

Evaluation of Audio-Visual Representation

Action Recognition Following prior works (Korbar, Tran, and Torresani 2018a; Han, Xie, and Zisserman 2019), to provide a fair comparison, we fine-tune the visual model on UCF-101 (Soomro, Zamir, and Shah 2012) and HMDB-51 (Kuehne et al. 2011) datasets, which consist of around 13K videos from 101 action classes and 7K videos from 51 action classes, respectively. Specifically, once we complete the audio-visual pre-training process, we use the learned parameters to initialize the visual model (S3D) but randomly initialize the last classification layer for action recognition. Additionally, to figure out the effect of temporal resolution and spatial resolution, we fine-tune our model with different input configurations.

Due to the large variability of experimental setups used in this task, like the backbone, pre-training dataset, and input spatial-temporal resolution, it is hard to conduct experiments with all the same settings. However, to present a set of meaningful comparisons, except classification accuracy, we also report other 5 factors (*e.g.,* "pre-train dataset", "Backbone", "Size", "Parameters" and "Flops") in Table 1. Notably, the overall results are grouped into four groups on UCF-101 and HMDB-51 datasets.

From the results in the table, we can make the following observations: 1) All the models pre-trained on larger unlabelled dataset have a remarkable boost in classification accuracy on small size dataset compared with the baseline model which fully train the model from scratch. This partly demonstrates that a meaningful pretext task can generate effective initialization for ConvNets to produce performance boost over randomly initialization. 2) Existing methods use a complex visual encoder, which is inefficient and computationally intractable. This is especially serious when deploying the model in a true scenario. 3) Compared with those self-supervision methods, our method yields better results. More specifically, compared with the pure video-based self-supervised methods, like (Xu et al. 2019), our method has at least 11.4% improvement on UCF-101 and 9.8% on HMDB-51 dataset. While compared with audio-visual-based self-

Method	Pre-train dataset	Backbone	Size	Parameters	Flops	UCF101	HMDB51
From scratch	-	S3D	16x224x224	8.3M	18.1G	52.7	39.2
Shuffle & Learn (Misra, Zitnick, and Hebert 2016)	UCF101/HMDB51	CaffeNet	1x227x227	58.3M	7.6G	50.2	18.1
Geometry (Gan et al. 2018)	UCF101/HMDB51	FlowNet	1x227x227	-	-	54.1	22.6
OPN (Lee et al. 2017)	UCF101/HMDB51	CaffeNet	1x227x227	58.3M	7.6G	56.3	23.8
ST order (Büchler, Brattoli, and Ommer 2018)	UCF101/HMDB51	CaffeNet	1x227x227	58.3M	7.6G	58.6	25.0
Cross & Learn (Sayed, Brattoli, and Ommer 2018)	UCF101/HMDB51	CaffeNet	1x227x227	58.3M	7.6G	58.7	27.2
CMC (Tian, Krishnan, and Isola 2019)	UCF101/HMDB51	CaffeNet	11x227x227	58.3M	83.6G	59.1	26.7
RotNet3D* (Jing et al. 2018)	Kinetics-400	3D-ResNet18	16x112x112	33.6M	8.5G	62.9	33.7
3D-ST-Puzzle (Kim, Cho, and Kweon 2019)	Kinetics-400	3D-ResNet18	16x112x112	33.6M	8.5G	63.9	33.7
Clip-order (Xu et al. 2019)	Kinetics-400	R(2+1)D-18	16x112x112	33.3M	8.3G	72.4	30.9
DPC (Han, Xie, and Zisserman 2019)	Kinetics-400	Custom 3D-ResNet	25x224x224	32.6M	85.9G	75.7	35.7
Multisensory (Owens and Efros 2018)	Kinetics-400	3D-ResNet18	64x224x224	33.6M	134.8G	82.1	-
CBT* (Sun et al. 2019)	Kinetics-400	S3D	16x112x112	8.3M	4.5G	79.5	44.6
L^3-Net (Arandjelovic and Zisserman 2017)	Kinetics-400	VGG-16	16x224x224	138.4M	113.6G	74.4	47.8
AVTS (Korbar, Tran, and Torresani 2018b)	Kinetics-400	MC3-18	25x224x224	11.7M	-	85.8	56.9
XDC* (Alwassel et al. 2019)	Kinetics-400	R(2+1)D-18	32x224x224	33.3M	67.4	86.8	47.1
SSCL-stage-I	Kinetics-400	S3D	16x112x112	8.3M	4.5G	81.4	47.7
SSCL-stage-II	Kinetics-400	S3D	16x112x112	8.3M	4.5G	82.6	49.9
SSCL-stage-II	Kinetics-400	S3D	16x224x224	8.3M	18.1G	84.3	54.1
SSCL-stage-II	Kinetics-400	S3D	32x224x224	8.3M	36.3G	**87.1**	**57.6**

Table 1: Overall comparison of our proposed SSCL method and the compared approaches on the action recognition benchmarks of UCF-101, and HMDB-51. Here * indicates the method is informally published on "arxiv.org" recently.

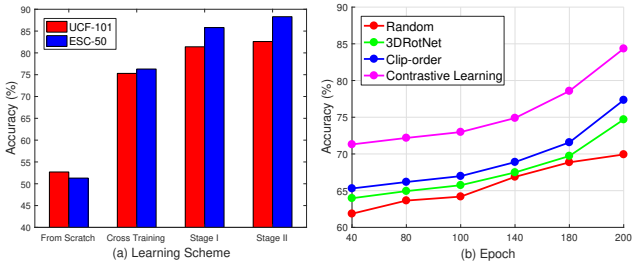

Figure 2: Effect of curriculum learning and pre-training strategy. Experiments are conducted on UCF-101 and ESC-50 datasets, where action and sound classification results are reported for curriculum learning analysis while action classification results are reported for pre-train strategies analysis.

supervised learning method, we obtain at least 0.3% gain on UCF-101 and 0.7% gain on HMDB-51 dataset. Additionally, from the model complexity group of the table, we only utilize a small video backbone but achieve a best performance. It indicates the effectiveness of our proposed SSCL method. 4) The accuracy gained from stage-I to stage-II proves the effectiveness of the training scheme of our curriculum learning and that cross-modal training provides a strong self-supervision signal due to the concurrency nature of acoustic and visual message. To this end, we speculate that the single modality chaotic that does not consider before might confuse the pair matching training process.

Sound Recognition To evaluate the audio-feature our method learned, for a fair comparison, linear probing protocol is applied for the task of audio recognition, which is also a routine in this literature. For this purpose, following previous works (Korbar, Tran, and Torresani 2018a; Alwassel et al. 2019), we fix the audio encoder (2D-ResNet10) except the last classification layer and test the audio representation quality on two established sound classification datasets: EDC-50 (Piczak 2015b) and DCASE (Stowell et al. 2015), which contains 2000 audio clips from 50 balanced environment sound classes and 100 audio clips from 10 balanced scene sound classes, respectively. Specifically, the input audios are processed with a randomly sampled wave within 1 second at a 24kHZ sampling rate, and then extracted as a spectrogram with this sampled wave of size 128x128 (time and frequency bands).

The overall results are summarized in Table 2. Like the table presented in action recognition, we also list the terms "Pre-train dataset" and "Backbone" in this table. Notably, the following observations can be seen: 1) Even with linear probing, the audio representations from the fixed filter can outperform fully training ConvNets with random initialization. 2) Similarly to visual case, audio representations extracted by our method outperform prior work. Specifically, the result of our SSCL is at least 5.7% higher than previous works on ESC-50. It indicates that SSCL helps the model to gain distinguish ability between the instances in each modality. Intuitively, the elaborative cross-modal knowledge transfer efficiently works and the audio-visual correspondence helps in generating better uni-modal representations.

Further Analysis

Analyses on Curriculum Learning In our SSCL method, the stage-I of the curriculum learning to the teacher network is based on the assumption that the noisy and irrelevant information between the audio and visual modalities may affect the transfer process. To prove this hypothesis, we explore whether the pre-training in stage-I helps in information transfer between audio and visual correspondence. We first study variants of the curriculum learning to understand the influence of within-modality self instance discrimination and cross-modal discrimination. Different strategies are designed to solve the objective defined in Eq. 3. We conduct experiments with directly optimizing Eq. 3 and different states of curriculum learning. To this end, we evaluate

Method	Pre-train dataset	Backbone	ESC-50 (%)	DCASE (%)
From scratch	-	2D-ResNet10	51.3	75.0
CovNet (Piczak 2015a)	ESC-50/DCASE	Custom-2 CNN	64.5	-
ConvRBM (Sailor, Agrawal, and Patil 2017)	ESC-50/DCASE	Custom-2 CNN	86.5	-
SoundNet (Aytar, Vondrick, and Torralba 2016)	Flickr-SoundNet	VGG	74.2	88.0
DMC (Hu, Nie, and Li 2019)	Flickr-SoundNet	VGG	82.6	-
L^3-Net (Arandjelovic and Zisserman 2017)	Kinetics-400	VGG	79.3	**93.0**
AVTS (Korbar, Tran, and Torresani 2018b)	Kinetics-400	VGG	76.7	91.0
XDC* (Alwassel et al. 2019)	Kinetics-400	2D-ResNet18	78.0	-
SSCL-stage-I	Kinetics-400	2D-ResNet10	85.8	91.0
SSCL-stage-II	Kinetics-400	2D-ResNet10	**88.3**	**93.0**

Table 2: Overall comparison of our proposed SSCL method and the compared approaches on the standard sound recognition benchmarks of ESC-50, and DCASE. Notably, $\star$ indicates the method is informally published on "arxiv.org" recently.

the learned representations of the above setting on UCF-101 and ESC-50 datasets and report the results in Figure 2 (a).

From the figure, we can obtain the following observations: (1) Directly optimizing audio and visual knowledge transfer is a bad choice due to the noise. (2) The two stages of knowledge transfer between audio and visual typically have better results. (3) The transfer process is completed after stage-II results in better representations than stage-I due to the knowledge transfer between audio and visual.

From the results, we can conclude that the lack of within-modality calibration to obtain a self-instance discriminative property is bad for transfer, because the good visual representations can not only reflect visual feature similarities but also weaken the influence of noisy pair features. Additionally, the stage-II that exchanges the role of student and teacher is effective in producing a better model.

Analyses on Video Pre-training Strategy To figure out the influence of different pretext tasks in stage-I of curriculum learning, we compare the other two widely used self-supervised methods, 3DRotNet and Clip-order, in video representation. Subsequently, we continue the stage-II in curriculum learning. Note that all the methods use the same setting (audio and visual encoder, training details, *etc*) only differ in the process they use pretext task, which means we re-implement the compared methods and combine it to our transfer scheme. We report the classification accuracy of those methods obtained on the UCF-101 dataset with the S3D encoder. In particular, the encoder is first self-trained according to each pretext task. We then fine-tune the model on UCF-101 dataset with the checkpoint weights saved at epoch $\{40, 80, 100, 140, 180, 200\}$.

The overall results are reported in Figure 2 (b). It can be observed that, as the pre-training process continues, higher accuracy can be obtained in downstream classification task. Notably, the contrastive learning used in our method achieves the best results partly due to its large scale negative samples and effective training scheme.

Qualitative Analysis on Audio-Visual Correspondence To explore whether the features of audio-visual can be grouped together, we show cross-modal retrieval results with a ranked similar value. Similar to work (Arandjelovic and Zisserman 2017), we use the sampled videos from the Kinetics dataset to conduct experiments here. We report results of the top-5 positive visual samples according to the query of sound as shown in Figure 3. It can be observed that the proposed method can correlate well the semantically similar acoustical and visual information and group together semantically related visual concepts. We own it to the self-discriminator which distinguishes itself from others in the same modality and cross-modal transfer process which groups similar visual and audio information.

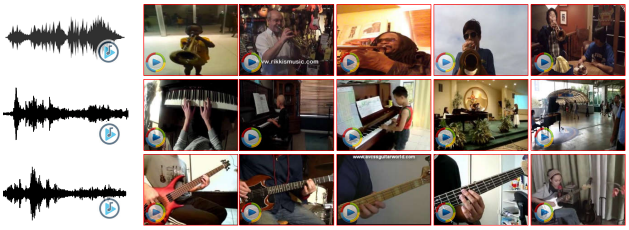

Figure 3: Audio and visual correspondence. For each query sound, we show five retrieval results in each row. The second one in each row is an optimal aligned audio and video pair from the same video.

Conclusion

In this paper, we aim to explore the close correlation between the acoustic signal and visual appearance in a self-supervised manner. We presented a cross-modal knowledge transfer framework with contrastive learning in the context of a teacher-student network paradigm to achieve that. In particular, a two-stage self-supervised curriculum learning scheme is proposed by solving the task of audio-visual correspondence learning. The rationale behind our method is that the knowledge shared between audio and visual modality serves as a supervisory signal. By using our framework, we can deploy the well-trained audio-visual model in practice to extract meaningful representations for a variety of downstream tasks, such as action recognition and audio recognition. To this end, our method extends the expressiveness of contrastive learning for audio-visual representation learning and provides a useful method for further research in this literature.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China (No. 2018AAA0102200); the National Natural Science Foundation of China under grants (No. 61976049, No. 61632007 and U20B2063); the Fundamental Research Funds for the Central Universities under Project (ZYGX2019Z015), the Sichuan Science and Technology Program, China (No. 2018GZDZX0032, 2019ZDZX0008, 2019YFG0533, 2019YFG0003) and National Science Foundation of Fujian Province (No. 2020J01084).

References

Alwassel, H.; Mahajan, D.; Torresani, L.; Ghanem, B.; and Tran, D. 2019. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667* .

Arandjelovic, R.; and Zisserman, A. 2017. Look, Listen and Learn. In *International Conference on Computer Vision*, 609–617.

Aytar, Y.; Vondrick, C.; and Torralba, A. 2016. SoundNet: Learning Sound Representations from Unlabeled Video. In *Advances in Neural Information Processing Systems*, 892–900.

Büchler, U.; Brattoli, B.; and Ommer, B. 2018. Improving Spatiotemporal Self-supervision by Deep Reinforcement Learning. In *European Conference on Computer Vision*, volume 11219, 797–814.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* .

Chung, S.; Chung, J. S.; and Kang, H. 2019. Perfect Match: Improved Cross-modal Embeddings for Audio-visual Synchronisation. In *International Conference on Acoustics, Speech and Signal Processing*, 3965–3969.

Gan, C.; Gong, B.; Liu, K.; Su, H.; and Guibas, L. J. 2018. Geometry guided convolutional neural networks for self-supervised video representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5589–5597.

Gan, C.; Huang, D.; Zhao, H.; Tenenbaum, J. B.; and Torralba, A. 2020. Music Gesture for Visual Sound Separation. In *Conference on Computer Vision and Pattern Recognition*, 10475–10484.

Gan, C.; Zhao, H.; Chen, P.; Cox, D. D.; and Torralba, A. 2019. Self-Supervised Moving Vehicle Tracking With Stereo Sound. In *International Conference on Computer Vision*, 7052–7061.

Gao, R.; Feris, R. S.; and Grauman, K. 2018. Learning to Separate Object Sounds by Watching Unlabeled Video. In *European Conference on Computer Vision*, volume 11207, 36–54.

Gao, R.; Oh, T.; Grauman, K.; and Torresani, L. 2020. Listen to Look: Action Recognition by Previewing Audio. In *Conference on Computer Vision and Pattern Recognition*, 10454–10464.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *International Conference on Learning Representations*.

Gupta, S.; Hoffman, J.; and Malik, J. 2016. Cross Modal Distillation for Supervision Transfer. In *Conference on Computer Vision and Pattern Recognition,*, 2827–2836.

Han, T.; Xie, W.; and Zisserman, A. 2019. Video Representation Learning by Dense Predictive Coding. In *International Conference on Computer Vision Workshops*, 1483–1492.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Conference on Computer Vision and Pattern Recognition*, 9726–9735.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hénaff, O. J.; Srinivas, A.; De Fauw, J.; Razavi, A.; Doersch, C.; Eslami, S.; and Oord, A. v. d. 2019. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272* .

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* .

Hu, D.; Nie, F.; and Li, X. 2019. Deep Multimodal Clustering for Unsupervised Audiovisual Learning. In *Conference on Computer Vision and Pattern Recognition*, 9248–9257.

Hu, D.; Wang, Z.; Xiong, H.; Wang, D.; Nie, F.; and Dou, D. 2020. Curriculum Audiovisual Learning. *arXiv preprint arXiv:2001.09414* .

Jing, L.; Yang, X.; Liu, J.; and Tian, Y. 2018. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387* .

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

Khosravan, N.; Ardeshir, S.; and Puri, R. 2019. On Attention Modules for Audio-Visual Synchronization. In *Conference on Computer Vision and Pattern Recognition Workshops, (CVPR Workshops)*, 25–28.

Kidron, E.; Schechner, Y. Y.; and Elad, M. 2005. Pixels that Sound. In *Conference on Computer Vision and Pattern Recognition*, 88–95.

Kim, D.; Cho, D.; and Kweon, I. S. 2019. Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles. In *Conference on Artificial Intelligence*, 8545–8552.

Korbar, B.; Tran, D.; and Torresani, L. 2018a. Co-training of audio and video representations from self-supervised temporal synchronization. *arXiv preprint arXiv:1807.00230* .

Korbar, B.; Tran, D.; and Torresani, L. 2018b. Cooperative Learning of Audio and Video Models from Self-Supervised

Synchronization. In *Advances in Neural Information Processing Systems*, 7774–7785.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T. A.; and Serre, T. 2011. HMDB: A large video database for human motion recognition. In *International Conference on Computer Vision*, 2556–2563.

Lee, H.; Huang, J.; Singh, M.; and Yang, M. 2017. Unsupervised Representation Learning by Sorting Sequences. In *International Conference on Computer Vision*, 667–676.

Misra, I.; Zitnick, C. L.; and Hebert, M. 2016. Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification. In *European Conference on Computer Vision*, volume 9905, 527–544.

Owens, A.; and Efros, A. A. 2018. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In *European Conference on Computer Vision*, volume 11210, 639–658.

Owens, A.; Wu, J.; McDermott, J. H.; Freeman, W. T.; and Torralba, A. 2018. Learning Sight from Sound: Ambient Sound Provides Supervision for Visual Learning. *Int. J. Comput. Vis.* 126(10): 1120–1137.

Park, D. S.; Chan, W.; Zhang, Y.; Chiu, C.; Zoph, B.; Cubuk, E. D.; and Le, Q. V. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Conference of the International Speech Communication Association*, 2613–2617.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.; and Póczos, B. 2019. Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities. In *Conference on Artificial Intelligence*, 6892–6899.

Piczak, K. J. 2015a. Environmental sound classification with convolutional neural networks. In *International Workshop on Machine Learning for Signal Processing*, 1–6.

Piczak, K. J. 2015b. ESC: Dataset for Environmental Sound Classification. In *Conference on Multimedia Conference*, 1015–1018.

Sailor, H. B.; Agrawal, D. M.; and Patil, H. A. 2017. Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification. In *Conference of the International Speech Communication Association*, 3107–3111.

Sayed, N.; Brattoli, B.; and Ommer, B. 2018. Cross and learn: Cross-modal self-supervision. In *German Conference on Pattern Recognition*, 228–243. Springer.

Shen, H. T.; Liu, L.; Yang, Y.; Xu, X.; Huang, Z.; Shen, F.; and Hong, R. 2020. Exploiting Subspace Relation in Semantic Labels for Cross-modal Hashing. *IEEE Transactions on Knowledge and Data Engineering* .

Shukla, A.; Petridis, S.; and Pantic, M. 2020. Learning Speech Representations from Raw Audio by Joint Audiovisual Self-Supervision. *International Conference on Machine Learning Workshop on Self-supervision in Audio and Speech* .

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* .

Stowell, D.; Giannoulis, D.; Benetos, E.; Lagrange, M.; and Plumbley, M. D. 2015. Detection and Classification of Acoustic Scenes and Events. *IEEE Trans. Multimedia* 17(10): 1733–1746.

Sun, C.; Baradel, F.; Murphy, K.; and Schmid, C. 2019. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743* .

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849* .

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Representation Distillation. In *International Conference on Learning Representations*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* .

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial Cross-Modal Retrieval. In *ACM on Multimedia Conference*, 154–162.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In *European Conference on Computer Vision*, volume 11219, 318–335.

Xu, D.; Xiao, J.; Zhao, Z.; Shao, J.; Xie, D.; and Zhuang, Y. 2019. Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. In *Conference on Computer Vision and Pattern Recognition*, 10334–10343.

Xu, G.; Liu, Z.; Li, X.; and Loy, C. C. 2020a. Knowledge Distillation Meets Self-Supervision. In *European Conference on Computer Vision*.

Xu, X.; Lu, H.; Song, J.; Yang, Y.; Shen, H. T.; and Li, X. 2020b. Ternary Adversarial Networks With Self-Supervision for Zero-Shot Cross-Modal Retrieval. *IEEE Trans. Cybern.* 50(6): 2400–2413.

Xu, X.; Wang, T.; Yang, Y.; Zuo, L.; Shen, F.; and Shen, H. T. 2020. Cross-Modal Attention With Semantic Consistence for Image–Text Matching. *IEEE Transactions on Neural Networks and Learning Systems* 31(12): 5412–5425.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful Image Colorization. In *European Conference on Computer Vision*, volume 9907, 649–666.

Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J. H.; and Torralba, A. 2018. The Sound of Pixels. In *European Conference on Computer Vision*, volume 11205, 587–604.