Audio-Visual Contrastive Learning with Temporal Self-Supervision

Simon Jenni,[1] **Alexander Black,** [2] **John Collomosse** [1,2]

[1] Adobe Research
[2] University of Surrey
jenni@adobe.com, alex.black@surrey.ac.uk, collomos@adobe.com

Abstract

We propose a self-supervised learning approach for videos that learns representations of both the RGB frames and the accompanying audio without human supervision. In contrast to images that capture the static scene appearance, videos also contain sound and temporal scene dynamics. To leverage the temporal and aural dimension inherent to videos, our method extends temporal self-supervision to the audio-visual setting and integrates it with multi-modal contrastive objectives. As temporal self-supervision, we pose playback speed and direction recognition in both modalities and propose intra- and inter-modal temporal ordering tasks. Furthermore, we design a novel contrastive objective in which the usual pairs are supplemented with additional sample-dependent positives and negatives sampled from the evolving feature space. In our model, we apply such losses among video clips and between videos and their temporally corresponding audio clips. We verify our model design in extensive ablation experiments and evaluate the video and audio representations in transfer experiments to action recognition and retrieval on UCF101 and HMBD51, audio classification on ESC50, and robust video fingerprinting on VGG-Sound, with state-of-the-art results.

Introduction

Videos provide a rich source of information for audio-visual learning. Besides static moments in time (single video frames), they also contain the scene dynamics (object motion) and often include the sounds of the environment and scene objects. It seems hopeless to learn general representations that capture this rich semantic information in videos, *i.e.*, their appearance, motions, and sounds from such high-dimensional data through sparse human supervision. Self-supervised learning (SSL) (**???**) has emerged as a viable alternative to supervised learning in recent years. Such methods might be better suited for general video representation learning since they are not constrained by the prohibitive cost of exhaustive human annotations on video. However, since most current self-supervised methods are tailored to static images, they might not effectively use videos' added temporal and aural dimensions. A self-supervised learning task that successfully integrates the static scene appearance and the aural and temporal features potentially results in a representation that better generalizes to downstream vision applications, such as action recognition, video retrieval, or robust video content fingerprinting.

Indeed, recent works that explored the aural and temporal dimensions of videos in isolation have demonstrated that they are both effective self-supervision signals. Several works (**???**) demonstrate that audio-visual contrastive learning often performs better than uni-modal contrastive learning (*i.e.*, using only the RGB frames). Likewise, temporal reasoning tasks (**???**) have demonstrated good transfer performance, especially for downstream tasks where motion is the main discerning factor (as opposed to static scene appearance).

In contrast, our work aims to leverage both sound and time as learning signals in a unified model architecture and training objective. To this end, we extend temporal self-supervision to the audio domain and propose cross-modal audio-visual temporal reasoning tasks. Concretely, we pose playback-speed and -direction recognition (**???**), as a pretext task for audio representation learning and propose temporal clip ordering as a task for both intra-modal (*e.g.*, audio-audio) and cross-modal (*e.g.*, audio-video) learning (see Figure 2). Furthermore, we introduce a model architecture and training objective for contrastive audio-visual learning that supplements these temporal learning tasks. Towards this goal, we carefully study how the inclusion and exclusion of different intra- and inter-modal contrastive objectives influences downstream performance. Our key findings for optimal audio-visual contrastive learning are 1. inclusion of video-video contrastive terms 2. temporally aligned cross-modal positives, and 3. exclusion of audio-audio contrastive terms (see Figure 1).

We further explore the design of the contrastive loss terms (**?**), *i.e.*, how to build positive and negative pairs for effective learning. In constructing our contrastive objective, we take inspiration from recent image-based methods (**??**) and extend the set of positive samples with nearest neighbors in the evolving feature space. Thus, besides standard augmented views for positive sampling, we consider nearest neighbors sampled from a queue of prior embeddings as additional positives. Notably, the neighborhood structure and sample weights are both calculated through cross-view similarity, *i.e.*, either through the feature space similarity to the augmented view (for intra-modal learning) or the temporally aligned sample from the other modality (for cross-modal learning). We also use this cross-view induced neighborhood structure to sample negative pairs in a sample-dependent

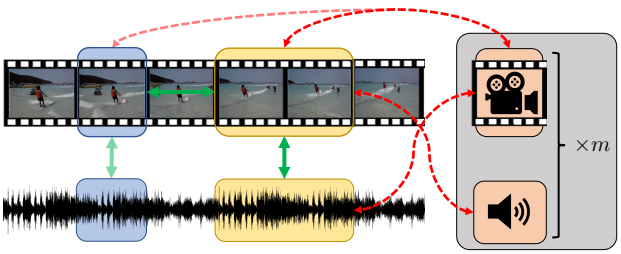

Figure 1: **Illustration of Contrastive Loss Terms in our Model.** We demonstrate the main contrastive pairs in our formulation given an example video clip (yellow box in the middle) and its corresponding audio clip. Positives (solid green arrows) are constructed from differently augmented video clips of the same training instance (blue box) and *temporally aligned* pairs of the corresponding video and audio clips. Negatives (dashed red arrows) stem from m other video and audio clips from the current mini-batch or a memory bank of prior embeddings (gray box on the right). Additional positives from the memory bank are omitted from the figure. Note that our formulation does not contain any contrastive terms among audio clips.

manner. This allows us to control the difficulty of the negative samples, *e.g.*, preventing ambiguous or confusing negatives resulting from duplicates or heavy class imbalance, while also preventing possible collapse through the absence of any negatives.

We verify our model design in extensive ablation experiments and compare it to prior works in established action recognition and retrieval benchmarks on UCF101 and HMDB51. We also evaluate the audio branch of our model for environmental sound classification on ESC50. Finally, we demonstrate the effectiveness of fusing the learned audio-visual features for downstream video classification on Kinetics-600 and VGG-Sound, and for robust video content retrieval under novel content manipulations for video fingerprinting (**??**) on VGG-Sound. We investigate video fingerprinting as a novel downstream application due to its growing importance given the ever-expanding scale of visual data online and the increasing threat and sophistication of malicious content manipulations.

Contributions. To summarize, we make the following contributions: 1) We introduce temporal self-supervision in the audio domain and the cross-modal setting; 2) We propose a contrastive loss design that extends the usual contrastive pairs with sample-dependent positives and negatives; 3) We explore various multi-modal contrastive model designs and demonstrate the importance of a) using temporally aligned positives for cross-modal terms and b) excluding audio-audio contrastive terms; 4) Finally, we demonstrate the quality of the learned audio-visual features in extensive transfer experiments to action recognition, video retrieval, audio classification, and a novel video fingerprinting benchmark.

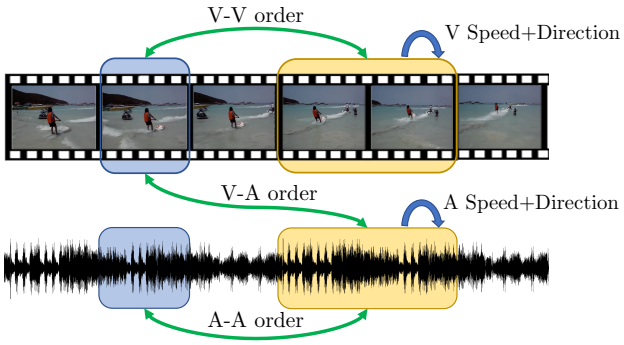

Figure 2: **Illustration of the Temporal Reasoning Tasks.** Besides contrastive terms, our model encompasses both per-clip classification tasks (blue arrows) about the playback-speed and -direction, and temporal ordering tasks (green arrows) which are performed both intra- and cross-modal (V: RGB frames, A: audio).

Prior Work

Contrastive Video Representation Learning. Contrastive learning is arguably the most popular self-supervised learning approach in computer vision today. These methods are typically based on the task of discriminating training instances up to strong data-augmentation (**??**), which was shown to be remarkably effective for unsupervised image representation learning (**??**) and has inspired a line of novel self-supervised methods (**????**). Recently, methods were proposed that extend the set of positive pairs with nearest neighbors in the learned embedding space (**??**). Our loss design similarly uses the evolving feature space to extend the set of contrastive pairs. In contrast, our loss design retains the exact match, contains multiple positives weighted based on cross-view similarity, and uses additional sample-dependent negatives.

Several recent works have explored contrastive learning on video. When dealing with video, the set of data augmentations can be extended with several temporal augmentations (*e.g.*, temporal crops). A natural extension is thus to add temporal augmentations to the set of data-augmentations that define the positive pairs for contrastive learning (**??**). Other works instead propose to learn to discriminate among temporally augmented clips (**??**), or learn to recognize the temporal input transformations in a multi-task approach (**??**). Our model combines contrastive learning among video clips with audio-visual contrastive and temporal self-supervised learning.

Temporal Self-Supervision. Classic self-supervised approaches were based on so-called pretext tasks. On images, popular examples are the ordering of image patches (**??**), the colorization of gray-scale images (**??**), or the classification of sets of image transformations (**???**). Pretext tasks that turned out particularly successful on video are based on recognizing temporal transformations. Some works explored the ordering of video frames (**????**) or whole video clips (**??**), others the classification of the playback direction (**?**), the playback speed (**???**), or general temporal warpings

(**?**). We also leverage temporal supervision and extend it to multi-modal audio-visual representation learning.

Audio-Visual Self-Supervised Learning. Another source of self-supervision on video can be found in the accompanying sound. Early works explored audio to learn single frame representations, *e.g.*, by predicting summary statistics of the sounds corresponding to a frame (**?**), or by recognizing if an audio snippet and image are temporally aligned (**??**). Similar to these image-based approaches (**?**) learned audio and video representations by recognizing when audio and video signals are synchronized. More recently, contrastive audio-visual learning for video achieved remarkable performance (**?**). For example, (**?**) performs clustering in one domain (*e.g.*, audio) and uses the resulting clusters as supervision for the other domain (*e.g.*, video). (**?**) demonstrate the effectiveness of cross-modal audio-visual contrastive learning and extend the set of positive samples within a modality with samples that show high cross-modal agreement. Other works even include language in audio-visual contrastive models (**??**). We instead focus on audio-visual learning and propose incorporating temporal supervision in both modalities.

Model

Let $\mathcal{D}_v = \{v_1, v_2, \ldots, v_N\}$ be a set of unlabelled training videos and let $\mathcal{D}_a = \{a_1, a_2, \ldots, a_N\}$ be their corresponding audio tracks. Our goal is to learn a video encoder F_v (a 3D-ConvNet) and an audio encoder F_a (a 2D-ConvNet) without human supervision. The inputs to the two networks are assumed to be of shape $v_i \in \mathbb{R}^{T\times H\times W\times C}$ and $a_i \in \mathbb{R}^{f\times t}$, where a_i is a spectogram representation of the audio track.

Temporal Input Augmentations. An essential component of modern SSL approaches is the set of data augmentations applied to the input. In contrastive learning, these input transformations define the set of learned invariances. They typically comprise color jittering and geometric transformations, like random resizing, cropping, and horizontal flipping. For our method, temporal transformations, *i.e.*, random temporal cropping and manipulations of the playback speed and direction, are particularly important. We will thus indicate the precise temporal manipulations with τ_r. Furthermore, we assume that τ_r has consistent behavior across modalities, *i.e.*, $\tau_r(v_j)$ and $\tau_i(a_r)$ represent the exact same moments in time for the audio and video domain.

Intra- and Inter-Modal Contrastive Learning

Our training objective comprises several predictive-contrastive loss terms. In general, we formulate these losses based on the two modalities involved and on the direction of the prediction, *e.g.*, indicating that the visual representation is being predicted from the audio. For the purpose of this discussion let $\nu_i^r = \psi_v(F_v(\tau_r(v_i)))$ denote the output of the video encoder followed by a projection MLP ψ_v for the input $\tau_r(v_i)$. Let similarly $\alpha_i^r = \psi_a(F_a(\tau_r(a_i)))$ be the feature vector for the corresponding audio track. Let further $\hat{\nu}_i^s$ be the feature of a different augmentation of the video v_i.

We illustrate the general form of the contrastive objective using the video-to-audio loss term, which is given by

$$\ell_{va}(\nu_i^r, \mathcal{P}_i^{va}, \mathcal{N}_i^{va}) = \sum_{p,w\in\mathcal{P}_i^{va}} -w \log\left(\frac{d(\phi_v(\nu_i^r), p)}{d(\phi_v(\nu_i^r), p) + \sum_{n\in\mathcal{N}_i^{va}} d(\phi_v(\nu_i^r), n)}\right), \tag{1}$$

where ϕ_v denotes a predictor MLP (following prior work (**?**)) and

$$d(x, y) := \exp\left(\frac{1}{\lambda}\frac{x^\intercal y}{\|x\|_2\|y\|_2}\right), \tag{2}$$

is a measure of the similarity between the feature representations of x and y, and $\lambda = 0.2$ is a temperature parameter. Note that we do not back-propagate through the second argument y in Eq. 2. In this general formulation, the set $\mathcal{P}_i$ defines the instance-dependent positive samples along with their weighting factor w, and $\mathcal{N}_i$ defines the negatives for contrastive learning.

Sources for Positive and Negative Contrastive Samples. We consider two sources for sampling the positive and negative pairs of the contrastive loss terms: 1. the set of examples in the mini-batch $\mathcal{B}$ at each iteration, and 2. a memory bank of prior feature embeddings. In our model, we maintain a memory bank Q_v (implemented as a FIFO queue) for the video domain and a corresponding Q_a for audio. Let $|Q_v| = |Q_a| = n_q$ be the size of the memory banks and let $\text{NN}_{j:k}(\nu, Q_v)$ denote the sequence $\{\eta_j, \ldots, \eta_k\}$ from the j-th to the k-th nearest-neighbor of ν in Q_v. For positive examples from the memory bank we further introduce a set of loss term weights $W_{1:k}(\nu) := \{\omega_1, \ldots, \omega_k\}$, where each weight is given by

$$\omega_j := \frac{d(\nu, \eta_j)}{\sum_{\eta_l\in\text{NN}_{1:k}(\nu_i, Q_v)} d(\nu, \eta_l)}, \tag{3}$$

thus weighting each nearest neighbor proportional to their similarity to ν. The memory banks are updated with the mean of the features from the two augmented views in each mini-batch, *i.e.*, $(\nu + \hat{\nu})/2$ in the case of Q_v.

We will now describe different instantiations of the contrastive losses and their positive and negative sample sets for the intra-modal and cross-modal objectives.

Visual-Visual Contrastive Term ℓ_{vv}. For a video feature vector ν_i^r in the case of video-video contrastive learning, we set $\mathcal{P}_i^{vv} = \{(\hat{\nu}_i^s, 1)\} \cup \text{NN}_{1:k}(\hat{\nu}_i^s, Q_v) \times W_{1:k}(\hat{\nu}_i^s)$, where $\text{NN}_{1:k}(\hat{\nu}_i^s, Q_v)$ is the set of the first k nearest neighbors of $\hat{\nu}_i^s$ extracted from Q_v. We set $k = 5$ in our experiments. The set of negatives is constructed as $\mathcal{N}_i^{vv} = \{\nu_j \in \mathcal{B} | j \neq i\} \cup \text{NN}_{q:q+m}(\hat{\nu}_i^s, Q_v)$ and contains all the video features not belonging to v_i that are in the current training mini-batch $\mathcal{B}$, as well as m additional negatives sampled from the memory queue as the q-th up to the $(q+m)$-th nearest neighbor of $\hat{\nu}_i^s$. By default we set $q = \frac{n_q}{2}$, thus starting from the neighbor in Q_v with median distance to $\hat{\nu}_i^s$ and set $m = 2048$.

Audio-Visual Contrastive Terms ℓ_{va} and ℓ_{av}. Since the terms ℓ_{va} and ℓ_{av} and the definition of their respective positive and negative sets is symmetric, we will restrict our illustration to the case of ℓ_{va}. Given a video feature vector ν_i^r,

we set $\mathcal{P}_i^{va} = \{(\alpha_i^r, 1)\} \cup \text{NN}_k(\alpha_i^r, Q_a) \times W_{1:k}(\alpha_i^r)$, where α_i^r is the feature of the corresponding audio clip with *identical temporal augmentation* (note the superscript). This is in contrast to the definition of ℓ_{vv} where positive pairs were not temporally aligned. As we will show in ablations, we found temporal alignment to be important for cross-modal contrastive learning. The set of negatives is defined as $\mathcal{N}_i^{va} = \{\nu_j \in \mathcal{B} | j \neq i\} \cup \{\alpha_j \in \mathcal{B} | j \neq i\} \cup \text{NN}_{q:q+m}(\alpha_i^r, Q_a)$, *i.e.*, we consider both other audio and other video feature vectors as negatives.

Multi-Modal Contrastive Objective. Our final contrastive objective is composed of the following intra- and inter-modal terms

$$\mathcal{L}_{\text{CRL}} = \mathop{\mathbb{E}}_{v_i, a_i} [\ell_{vv}(\nu_i^r, \mathcal{P}_i^{vv}, \mathcal{N}_i^{vv}) + \ell_{va}(\nu_i^r, \mathcal{P}_i^{va}, \mathcal{N}_i^{va}) + \ell_{av}(\alpha_i^r, \mathcal{P}_i^{av}, \mathcal{N}_i^{av})]. \quad (4)$$

Note that our final model does not contain an audio-audio contrastive term. Indeed, we find that including such a term analogous to ℓ_{vv} hurts the final feature performance in transfer experiments (see ablations in Table 3). An illustration of the intra- and inter-modal terms is given in Figrue 1.

Temporal Self-Supervision for Video and Audio

Aside from learning from the correspondence between audio and video as proposed above, we also want to promote the learning of temporal features in both domains through self-supervised temporal reasoning tasks. These temporal pretext tasks can be categorized into unitary intra-modal tasks and pairwise intra- and cross-modal objectives (see Figure 2).

Intra-Modal Speed and Direction Classification. To capture short-term temporal video features we leverage the classification of temporal transformations as SSL objectives (**?**). Concretely, we train the model to predict whether videos are played forward or backward and at which playback speed. The direction classification is a simple binary classification task per clip, and either direction is equally likely during training. The speed classification is posed as a classification task among 4 speed classes ($1\times$, $2\times$, $4\times$, and $8\times$ speedup). The speed manipulations are implemented via temporal subsampling, and all the speed classes are equally likely during pre-training.

We propose to leverage such temporal supervision in the audio domain in this work. We apply the temporal transformations to the 1D raw audio signal (analogous to the video domain), *i.e.*, we subsample the signal for speed manipulations and reverse its direction before computing the spectrogram. In experiments, we also investigate an alternative approach where we perform the temporal transformations in the audio spectrogram (thus not manipulating the frequency). Interestingly, we found that transforming the raw audio waveform is much more effective, even when accounting for processing artifacts in manipulating the spectrogram (see ablations in Table 2).

Intra- and Inter-Modal Temporal Ordering. To capture the longer-term dynamics of videos we propose to also perform temporal learning tasks at the clip level by predicting the order of two video clips. Besides performing such temporal ordering solely on video (**???**), we extend it to temporal ordering of the audio tracks and cross-modal audio-visual temporal ordering. Concretely, we pose the three-way classification of two temporal signals into 1. correctly ordered, 2. overlapping, and 3. wrongly ordered. This task is implemented by concatenating the representations of the two time-signals along the channel dimension and feeding it through a classifier, *e.g.*, $\phi_{va}([F_v(v_i), F_a(a_i)])$ for video-audio ordering. Likewise, we introduce classifiers ϕ_{vv}, ϕ_{av}, and ϕ_{aa} for video-video, audio-video, and audio-audio temporal ordering.

Finally, we jointly optimize the network weights of the audio and video branch on the combination of the temporal and contrastive objectives. Concretely, let $\mathcal{L}_{\text{TEMP}} = \mathcal{L}_{speed} + \mathcal{L}_{direction} + \mathcal{L}_{order}$ be the sum of all the losses for the above temporal reasoning tasks. The final objective is then given by

$$\mathcal{L}_{\text{SSL}} = \mathcal{L}_{\text{CRL}} + \lambda \mathcal{L}_{\text{TEMP}}, \quad (5)$$

where we set $\lambda = 0.5$.

Implementation Details

For our video encoder F_v we consider variants of the popular 3D-ConvNet architectures R3D (**?**) and R(2+1)D (**?**). If not specified otherwise, input video clips are assumed to contain 16 frames of resolution 112×112 for R(2+1)D, 128×128 for R3D-18, and 224×224 for R3D-34. Our audio encoder F_a is based on a standard ResNet-34 (**?**) architecture in all experiments. Input spectrograms to the audio encoder are resized to 224×224.

We train the models using the AdamW optimizer (**?**) with a weight decay set to 10^{-4}. The learning rate follows a cosine annealing schedule (**?**) with a maximum learning rate of $3 \cdot 10^{-4}$ and linear warm-up in the first training epoch. By default, we train all the models with a batch size of 256.

Besides the temporal input transformations described above (*i.e.*, playback speed+direction changes and temporal cropping), we use the typical data augmentation recipe for contrastive methods, *i.e.*, horizontal flipping, color-jittering, and random spatial cropping. We do not apply any augmentations beyond the temporal ones for audio.

The projection MLPs ψ contain two hidden layers of size 1024 and output feature embeddings of size 256. The prediction MLPs ϕ contain a single hidden layer with a hidden dimension of 1024. We apply synchronized batch norm in both MLPs (including the output of ψ) following prior work (**?**). The classification heads for the temporal self-supervision tasks follow a similar design to ψ, except that no batch norm is applied to the output in this case.

To evaluate models in transfer experiments, we average predictions of multiple temporal and spatial crops. Likewise, the features for linear probes and nearest-neighbor retrieval are obtained by averaging multiple crops and standardizing the resulting features using the training set statistics.

Experiments

Datasets. As a pre-training dataset we use Kinetics (**?**) in most of our experiments. The dataset contains around 350K training videos categorized into 600 human action

Table 1: **Contrastive Loss Design.** We explore different configurations of the contrastive loss formulation in Eq. 1 in combination with temporal SSL when applied to video-video learning (no audio is being used). We report nearest-neighbor classifier accuracy on UCF101 and HMDB51 and recall @1 for robust video fingerprinting on VGG-Sound.

Experiment	UCF101 1-NN	HMDB51 1-NN	AugVGG-C R@1
(a) w/o Q_v positives	61.5	32.5	**65.5**
(b) w/o Q_v negatives	<u>63.9</u>	34.0	<u>65.1</u>
(c) hard negatives	63.5	33.3	**65.5**
(d) easy negatives	62.9	<u>34.8</u>	<u>65.1</u>
(e) uniform ω_j	63.7	33.1	64.1
Baseline	**65.3**	**35.3**	**65.5**
(f) NNCLR	64.8	34.2	62.8
(g) SimCLR	53.9	29.2	61.1
(h) SimSiam	62.8	34.0	60.9

classes. For transfer experiments we consider UCF101 (**?**) and HMDB51 (**?**) which are significantly smaller datasets with human action annotations. We use these datasets to evaluate the transfer performance of the video branch, both via fine-tuning to action recognition and as fixed feature extractors for video retrieval. We evaluate the audio branch of our model on ESC50 (**?**) in terms of environmental audio classification.

Augmented VGG-Sound. Finally, we use the test set of VGG-Sound (**?**) to evaluate both branches in terms of their robustness to heavy content manipulation for fingerprinting applications. Concretely we generate the following four augmented versions of the dataset by applying different types of audio and video transformations (examples in parenthesis):

1. AugVGG-IP - "In-Place" manipulations (V: noise, blur, pixelization, emoji overlay; A: noise, clicks).

2. AugVGG-S - "Spatial" transformations (V: cropping, padding, rotation; A: pitch shift, reverb, freq. filter).

3. AugVGG-T - "Time" transforms (V+A: speed, crops).

4. AugVGG-C - "Combined" (one of each type above).

We use the AugLy library for the dataset creation (**?**). For fingerprinting evaluations, we report recall at k for these datasets where queries stem from AugVGG-x and retrievals are computed on the clean test set.

Ablations

We perform extensive ablation experiments to investigate the influence of the contrastive loss function design, the various temporal self-supervision signals for audio representation learning, and our combined audio-visual model.

On the Design of the Contrastive Loss. We perform experiments with different variants of the general contrastive objective in Equation 1 and compare it to some popular existing baselines. For faster experimentation, we perform these experiments on video only (we do not use the audio channel here) and pre-train the networks for 40 epochs. We use an R3D-18 network architecture and perform the temporal reasoning tasks among video clips in the experiments. We compare the following variants and report results in Table 1:

(a)-(b) Positives and negatives from the memory bank. In this case, we remove the nearest neighbors from the memory bank as additional positives (a) or remove the negative sampling from Q_v (b). We observe that both positives and negatives from Q_v demonstrate clear benefits, while the positives provide more significant improvements, especially in action retrieval performance.

(d)-(e) Difficulty of negatives. Instead of sampling negatives starting from the median of nearest neighbors in the memory bank, we start at the 90th percentile for hard negatives in (c) and at the 20th percentile for easy negatives (d). Both variants lead to inferior action retrieval performance, and easy negatives hurt fingerprinting.

(f) Equal weighting of positives. Instead of the cross-view similarity-based weighting of the positives, all five positive examples contribute equally to the loss in this case. We observe a drop in the fingerprinting retrieval especially, possibly due to decreased importance of the exact match in the loss. This case is similar to the approach in (**?**).

(g)-(i) Prior approaches. We replace our proposed loss with existing prior approaches. NNCLR (**?**) replaces the embedding of one view with its nearest neighbor in the memory bank. While this leads to good performance in action retrieval, the performance for fingerprinting suffers. We hypothesize that the lack of the exact match and the lack of additional negatives are the main reason. Key differences to SimCLR (**?**) are 1. lack of nearest neighbors, 2. lack of predictor MLP, 3. gradient back-propagation through both views. SimCLR requires much larger mini-batches to perform well, which is prohibitive on video. Finally, SimSiam (**?**) lacks any negative examples but is otherwise identical to (a). We can again observe the importance of explicit negatives for the fingerprinting use case.

Table 2: **Temporal Self-Supervision for Audio Feature Learning.** We explore how the different temporal self-supervision signals impact the audio representation performance for downstream audio classification on ESC50 and audio fingerprinting on VGG-Sound. The audio encoder is pre-trained with temporal supervision and audio-audio contrastive learning (no RGB frames were used).

Ablation	ESC50 Linear	ESC50 1-NN	AugVGG-C R@1
(a) w/o speed	80.4	58.4	21.1
(b) w/o direction	79.0	56.7	<u>21.8</u>
(c) w/o order	<u>80.6</u>	<u>58.5</u>	21.5
(d) spect.-resize	71.0	50.8	19.3
(e) + rand. STFT-step	76.5	53.3	21.5
Baseline	**82.2**	**61.0**	**21.9**

The Benefits of Temporal Self-Supervision for Audio. We performed ablation experiments to demonstrate the different temporal learning tasks' effect on audio feature performance. We only train the audio branch in these experiments and combine the temporal tasks with an audio-audio contrastive term. Networks were again trained for 40 epochs on

Table 3: **Audio-Visual Model Ablations.** We perform ablation experiments to demonstrate the influence of the different self-supervised learning signals in our approach (first block) and various implementation details (second block). The video encoder is evaluated in transfer to action recognition on UCF101 and HMDB51, and the audio encoder for classification on ESC50. The fused audio-video feature is used for fingerprinting on VGG-Sound.

Experiment	UCF101 1-NN	HMDB51 1-NN	ESC50 1-NN	AugVGG-C R@1
(a) w/o A-V CLR	61.0	32.2	62.9	73.8
(b) w/o V-V CLR	61.0	33.8	67.3	69.6
(c) w/ A-A CLR	<u>69.1</u>	37.4	68.3	78.1
(d) w/o temp.-SSL	67.5	37.4	67.4	78.6
(e) unaligned A-V	68.4	37.2	68.9	78.8
(f) shared Q	68.8	<u>38.5</u>	**69.4**	**79.3**
Baseline	**70.7**	**40.5**	<u>69.0</u>	78.1

Kinetics. In Table 2 (a)-(c), we report the performance of models where each of the three temporal supervision signals is removed. We can observe that each task significantly benefits feature performance, especially in downstream audio recognition tasks. In ablation (d)-(e), the temporal speed transformations are realized by resizing the audio spectrogram instead of subsampling the raw audio signal. We observe clear performance degradations in these cases, even when randomizing the frame step of the STFT, which could prevent some possible shortcuts due to resizing artifacts.

Combined Contrastive and Temporal Audio-Visual Learning. Finally, we validate our combined audio-visual model through experiments demonstrating the importance of the inclusion (or exclusion) of the different contrastive and temporal objectives and ablate model design variations. In this set of experiments, we use an R(2+1)D-18 architecture for the video encoder, and we again train the model for 40 epochs. Table 3 shows the results of the following experiments:

(a)-(d) Training Objectives: We show the influence of the different contrastive intra- and inter-modal objectives in (a)-(c) and the addition of the temporal reasoning tasks in (d). We observe that the cross-modal term brings the most benefit, followed by including the intra-video term. Interestingly, the exclusion of the intra-audio term performs better in all cases. Finally, note how adding temporal self-supervision to the contrastive objectives provides significant gains across the board.

(e)-(f) Implementation Details: We further illustrate the importance of using temporally aligned positives in the cross-modal contrastive term in (e). We believe that the model can leverage the temporal audio-visual correspondence to better associate scene events with their sounds. Finally, in (f), we use only a single memory bank which we feed with the averages of the features from both modalities. Interestingly, this outperforms separate memory banks for fingerprinting and audio recognition.

Comparison to Prior Work on Video SSL

We compare against prior self-supervised video representation learning methods in transfer learning experiments for action recognition and retrieval on UCF101 and HMDB51. We train and evaluate two different video encoders in these comparisons: 1. a smaller-scale experiment with an R(2+1)D-18 trained at 112×112 and 2. a larger-scale experiment with an R3D-34 trained at 224×224 resolution.

Transfer to Action Recognition and Audio Classification. We compare on UCF101 and HMDB51 action recognition and ESC50 audio classification in Table 4, both with full fine-tuning and linear probes when available. A fair comparison to and among prior works is difficult due to significant differences in pre-training datasets, network architectures, input configurations, and training duration. We indicate some of these factors that are known to impact performance in the table. While there are prior works (**??**) reporting comparable performance in some tasks, they either use larger architectures, larger pre-training datasets, train for longer, or a combination of those. Our method is more efficient in comparison while still achieving state-of-the-art performance. Notably, when comparing the most common setting using R(2+1)D-18 trained on Kinetics-400, we outperform the best prior results by +3.1%, +9.0%, and +5.7% on UCF101, HMDB51, and ESC-50 respectively.

Video Retrieval Performance. We compare to the prior state-of-the-art approaches TCLR (**?**), GDT (**?**), Robust-xID (**?**), and TE-CVRL (**?**) in video retrieval benchmarks on UCF101 and HMDB51 in Table 5. Queries stem from the test set, and retrievals are computed on the training set of the respective dataset. A retrieval is assumed correct when the class of query and retrieval agree. We report recall at k for different nearest neighbors. Our model outperforms prior methods by a considerable margin.

Video Fingerprinting Performance on AugVGG. Finally, we report video retrieval performance under video manipulations in Figure 3. We report recall at k for all four datasets and three models: 1. fused audio and video features, 2. video-only, and 3. audio-only. The fused embedding (concatenation of audio and video features) performs best in all cases, followed by the video model. Surprisingly, AugVGG-IP with in-place augmentations is most difficult, while performance on AugVGG-S and AugVGG-T is close to perfect.

Audio-Visual Feature Fusion. We explore the fusion of the aural and visual features learned through our approach for downstream video understanding tasks. We compare linear probe accuracy for audio, video, and fused features learned on VGG-Sound and Kinetics-600 in Table 6. Interestingly, combining both modalities improves not only the audio-focused VGG-Sound benchmark but also the appearance-focused classification task on Kinetics-600.

Conclusions

We introduced a novel method to learn video and audio representations by exploiting temporal and audio-visual self-supervision. To learn temporal features, our model learns through time-related pretext tasks, which we extend to the audio domain and the cross-modal setting. We propose a

Table 4: **Action Recognition on UCF101 and HMDB51 and Audio Classification on ESC50.** We report action recognition accuracy after full fine-tuning and linear probe evaluation. We indicate the pre-training dataset, resolution, the number of frames, iterations (or epochs in brackets), and pre-training data modalities (V=RGB, A=audio).

							UCF101		HMDB51		ESC50
Method	**Dataset**	**Res.**	**Frames**	**It. [Ep.]**	**Network**	**Mod.**	FT	Lin.	FT	Lin.	Lin.
TE-CVRL (?)	K400	112	16	[200]	R(2+1)D-18	V	88.2		62.2		
CVRL (?)	K600	224	32	[800]	R3D-50	V	93.4	90.6	68.0	59.7	
MMV (?)	AS	224	32	500K	R(2+1)D-18	V+A	91.5	83.9	70.1	60.0	
BraVe (?)	AS	224	32	620K	R(2+1)D-18	V+A	**93.6**	90.0	70.8	63.6	
AVTS (?)	K400	224	25	[90]	MC3	V+A	85.8		56.9		76.7
XDC (?)	K400	224	32	900K	R(2+1)D-18	V+A	84.2		47.1		78.5
GDT (?)	K400	112	32	[200]	R(2+1)D-18	V+A	88.7		57.8		78.6
AVID (?)	K400	224	32	[400]	R(2+1)D-18	V+A	87.5		60.8		79.1
Ours	VGG-S	112	16	160K [240]	R(2+1)D-18	V+A	90.9	86.8	70.2	55.9	**87.9**
Ours	K400	112	16	200K [240]	R(2+1)D-18	V+A	91.8	88.0	71.2	58.2	84.8
Ours	K600	112	16	200K [150]	R(2+1)D-18	V+A	92.2	90.3	72.2	62.6	86.4
Ours	K600	224	16	400K [300]	R3D-34	V+A	**93.6**	**91.8**	**74.6**	**65.8**	85.5

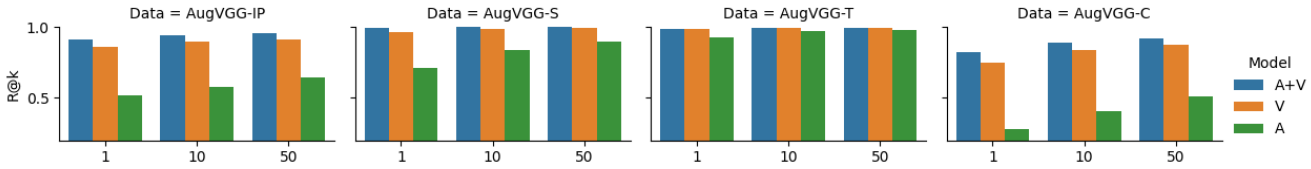

Figure 3: **Video Fingerprinting Performance.** We report instance retrieval performance under video content manipulation on the different AugVGG variants. We show results using a video only (V), audio only (A), and a joint audio-visual model (A+V).

Table 5: **Video Retrieval on UCF101 and HMDB51.** We report recall at k (R@k) for k-NN video retrieval. All methods use a R(2+1)D-18 network.

	UCF101			HMDB51		
Method	**R@1**	**R@5**	**R@20**	**R@1**	**R@5**	**R@20**
TCLR	56.9	72.2	84.6	24.1	45.8	75.3
GDT	57.4	73.4	88.1	25.4	51.4	75.0
Robust-xID	60.9	79.4	90.8	30.8	55.8	79.7
TE-CVRL	64.2	81.1	92.6	33.1	60.8	84.1
Ours (R(2+1)D-18)	**80.6**	**90.4**	**96.4**	**44.9**	**70.4**	**87.6**
Ours (R3D-34)	**85.2**	**93.0**	**97.3**	**51.3**	**74.3**	**91.4**

Table 6: **Modality Fusion.** We explore the fusion of our audio-visual features for downstream video classification.

Modalities	**VGG-Sound**	**K600**
Audio	39.1	15.7
Video	39.7	56.8
Audio+Video	**53.9**	**58.4**

novel contrastive loss design and a model with both intra- and cross-modal contrastive objectives to learn from the audio-visual correspondence in videos. Experiments demonstrate that representations that integrate both temporal and aural features achieve state-of-the-art video classification and retrieval performance.

Dolorem ea laudantium libero, impedit excepturi voluptatum maxime cupiditate illum non harum maiores atque, veniam corrupti fuga fugit excepturi, odit amet iusto fuga neque eaque autem ex rem veritatis blanditiis expedita, repellat aut quisquam?Numquam officia vero dignissimos, perspiciatis eos perferendis molestiae libero eaque fugit impedit architecto quas magni, veniam cum enim nulla dolor reiciendis eius asperiores doloremque, odit accusamus ex nostrum aspernatur earum alias sequi, amet eos blanditiis magnam animi libero quidem sunt molestiae vero?Voluptatem voluptatibus ratione a soluta vitae harum voluptatum quod, nesciunt nihil id nobis incidunt temporibus saepe facere nemo quo esse voluptate?Saepe iure eum architecto libero error aspernatur exercitationem eligendi animi quod, quia dolorum mollitia error porro autem enim, ratione omnis voluptatibus, ipsa error doloribus ipsum assumenda distinctio numquam libero.Doloribus nihil ducimus sint sapiente repellendus, consequatur iste facere voluptates, magni neque est molestiae aut alias porro praesentium, ea suscipit sequi, cumque provident vero veritatis deserunt repellat reprehenderit numquam possimus consequatur quod officiis.Soluta odit a deleniti unde cumque ab

enim inventore quibusdam, animi dolore ratione facere odio dolorum enim eligendi blanditiis accusamus libero, eveniet vel laudantium quod culpa odio, non