Audio-Visual Instance Discrimination with Cross-Modal Agreement

Pedro Morgado* UC San Diego
Nuno Vasconcelos UC San Diego
Ishan Misra Facebook AI Research

Abstract

We present a self-supervised learning approach to learn audio-visual representations from video and audio. Our method uses contrastive learning for cross-modal discrimination of video from audio and vice-versa. We show that optimizing for cross-modal discrimination, rather than within-modal discrimination, is important to learn good representations from video and audio. With this simple but powerful insight, our method achieves highly competitive performance when finetuned on action recognition tasks. Furthermore, while recent work in contrastive learning defines positive and negative samples as individual instances, we generalize this definition by exploring cross-modal agreement. We group together multiple instances as positives by measuring their similarity in both the video and audio feature spaces. Cross-modal agreement creates better positive and negative sets, which allows us to calibrate visual similarities by seeking within-modal discrimination of positive instances, and achieve significant gains on downstream tasks.

1. Introduction

Imagine the sound of waves. This sound can evoke the memory of many scenes - a beach, a pond, a river, *etc*. A single sound serves as a bridge to connect multiple sceneries. It can group visual scenes that 'go together', and set apart the ones that do not. We leverage this property of freely occurring audio to learn video representations in a self-supervised manner.

A common technique [2, 41, 62, 63] is to setup a verification task that requires predicting if an input pair of video and audio is 'correct' or not. A correct pair is an 'in-sync' video and audio and an incorrect pair can be constructed by using 'out-of-sync' audio [41] or audio from a different video [2]. However, a task that uses a *single* pair at a time misses a key opportunity to reason about the data distribution at large.

In our work, we propose a contrastive learning framework to learn cross-modal representations in a self-supervised manner by contrasting video representations against *multiple* audios at once (and vice versa). We leverage recent advances [28, 61, 80, 86] in contrastive learning to setup a Audio-Visual Instance Discrimination (AVID) task that learns a cross-modal similarity metric by grouping video and audio *instances* that co-occur. We show that the cross-modal discrimination task, *i.e*., predicting which audio matches a video, is more powerful than the within-modal discrimination task, predicting which video clips are from the same video. With this insight, our technique learns powerful visual representations that improve upon prior self-supervised methods on action recognition benchmarks like UCF-101 [76] and HMDB-51 [42].

We further identify important limitations of the AVID task and propose improvements that allow us to 1) reason about *multiple* instances and 2) optimize for visual similarity rather than just cross-modal similarity. We use Cross-Modal Agreement (CMA) to group together videos with high similarity in video and audio spaces. This grouping allows us to directly relate multiple videos as being semantically similar, and thus directly optimize for visual similarity in addition to cross-modal similarity. We show that CMA can identify semantically related videos, and that optimizing visual similarity among related videos significantly improves the learned visual representations. Specifically, CMA is shown to improve upon AVID on action recognition tasks such Kinetics [83], UCF-101 [76] and HMDB-51 [42] under both linear probing and full fine-tuning evaluation protocols.

2. Related work

Self-supervised learning is a well studied problem [13, 47, 52, 55, 60, 72]. Self-supervised methods often try to reconstruct the input data or impose constraints on the representation, such as sparsity [48, 59, 60], noise [82] or invariance [8, 10, 11, 17, 28, 35, 53, 69] to learn a useful and transferable feature representation. An emerging area of research uses the structural or domain-specific properties of visual data to algorithmically define 'pretext tasks'. Pretext tasks are generally not useful by themselves and are used as a proxy to learn semantic representations. They can use the spatial structure in images [16, 25, 57, 89], color [14, 45, 46, 90], temporal information in videos [18, 20, 29, 37, 49, 54, 58, 64, 85] among

*Work done while interning at Facebook AI Research.

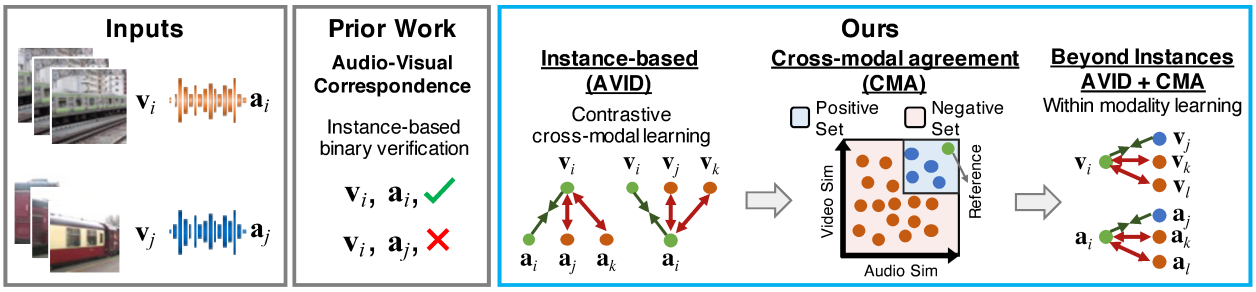


Figure 1: Popular audio-video self-supervised methods can be interpreted as 'instance-based' as they learn to align video and audio *instances* by solving a binary verification problem. We propose AVID to learn cross-modal representations that align video and audio instances in a contrastive learning framework. However, AVID does not optimize for visual similarity. We calibrate AVID by formulating CMA. CMA finds groups of videos that are similar in both video and audio space which enables us to directly optimize representations for visual (within modality) similarity by using these groups.

other sources of 'self' or naturally available supervision. We propose an unsupervised learning technique that leverages the naturally available signal in video and audio alignment.

Representation Learning using Audio. Self-supervised learning can also make use of multiple modalities, rather than the visual data alone. As pointed out in [13, 38], co-occurring modalities such as audio can help learn powerful representations. For example, audio self-supervision has shown to be useful for sound source localization and separation [3, 21–23, 74, 92, 93], lip-speech synchronization [12] and visual representation learning [2, 41, 62] and audio spatialization [56].

Audio-Visual Correspondence (AVC) is a standard task [2, 3, 41, 62] used in audio-video cross-modal learning. This task tries to align the visual and audio inputs by solving a binary classification problem. However, most methods use only a single video and single audio at a time for learning. Thus, the model must reason about the distribution over multiple samples implicitly. In our work, we use a contrastive loss [28, 61, 80, 86] that opposes a large number of samples simultaneously. We show in §5 that our method performs better than recent methods that use AVC.

Contrastive Learning techniques use a contrastive loss [28] to learn representations either by predicting parts of the data [32, 33, 61], or discriminating between individual training instances [17, 19, 31, 34, 53, 86, 88, 94]. Contrastive learning has also been used for learning representations from video alone [29, 75]. Tian *et al*. [80] also use a contrastive approach, but propose to learn with a cross-modal objective applied to images and depth, video and flow. In contrast, our method learns visual representations using audio as cross-modal targets. Compared to [80], we present a new insight for audio-visual learning that optimizing cross-modal similarity is more beneficial than within-modal similarity. We also identify important limitations of cross-modal discrimination and present an approach that goes beyond instance discrimination by modeling Cross-Modal Agreement. This identifies groups of related videos and allows us to optimize for within-modal similarity between related videos. The concurrently proposed [1] uses alternating optimization to find clusters in visual and audio feature spaces, *independently* and uses them to improve cross-modal features. While our CMA method bears a resemblance to theirs, we do not use alternating optimization and use *agreements* between the visual and audio representations to directly improve visual similarity rather than only cross-modal similarity. Finally, similar to our work, the concurrently proposed [30] also uses co-occurring modalities (optical flow and RGB) to expand the positive set. However, instead of mining positives based on an agreement between both modalities, [30] relies on the opposite modality alone.

Multi-view Learning. Multi-view learning aims to find common representations from multiple views of the same phenomenon, and has been widely used to provide learning signals in unsupervised and semi-supervised applications. Classical approaches can be broadly categorized in co-training procedures [6, 7, 30, 43, 50, 67, 84] that maximize the mutual agreement between views, multiple kernel learning procedures [5, 40, 44] which use kernels to model different views, and subspace learning procedures [15, 68] which seek to find the latent space that generates all views of the data.

Multi-view data is an effective source of supervision for self-supervised representation learning. Examples include the motion and appearance of a video [30, 80], depth and appearance [36, 91], luminance and chrominance of an image [80, 91], or as in our work sound and video [2, 4, 12, 63].

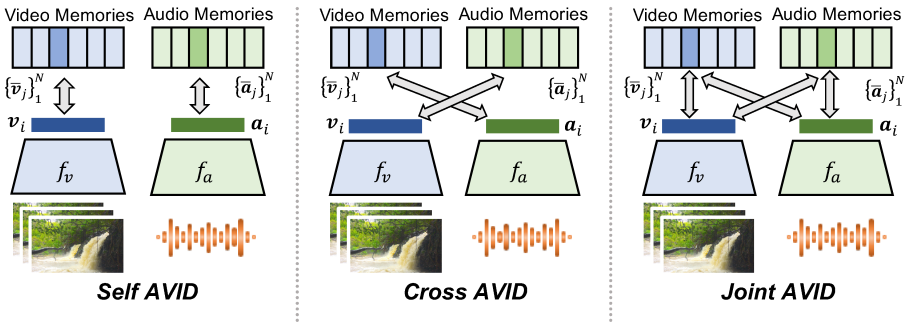


Figure 2: Variants of the AVID task. Instance discrimination can be accomplished contrasting representations within the same modality (Self-AVID), across modalities (Cross-AVID) or a mixture of the two (Joint-AVID).

3. Audio-Visual Instance Discrimination

We learn visual representations in a self-supervised manner from unconstrained video and audio by building upon recent advances in instance discrimination [17, 51, 80, 86] and contrastive learning [27, 28, 61].

3.1. Goal and Intuition.

Consider a dataset of N samples (instances) $\mathcal{S} = \{s_i\}_{i=1}^N$ where each instance s_i is a video s_i^v with a corresponding audio s_i^a. The goal of Audio-Visual Instance Discrimination (AVID) is to learn visual and audio representations $(\mathbf{v}_i, \mathbf{a}_i)$ from the training instances s_i. The learned representations are optimized for 'instance discrimination' [17, 51, 86], *i.e.*, must be discriminative of s_i itself as opposed to other instances s_j in the training set. Prior work [17, 86] shows that such a discriminative objective among instances learns semantic representations that capture similarities between the instances.

To accomplish this, two neural networks extract unit norm feature vectors $\mathbf{v}_i = f_v(s_i^v)$ and $\mathbf{a}_i = f_a(s_i^a)$ from the video and audio independently. Slow moving (exponential moving average) representations for both video and audio features $\{(\bar{\mathbf{v}}_i, \bar{\mathbf{a}}_i)\}_{i=1}^N$ are maintained as 'memory features' and used as targets for contrastive learning. The AVID task learns representations $(\mathbf{v}_i, \mathbf{a}_i)$ that are more similar to the memory features of the instance $(\bar{\mathbf{v}}_i, \bar{\mathbf{a}}_i)$ as opposed to memory features of other instances $(\bar{\mathbf{v}}_j, \bar{\mathbf{a}}_j)$, $j \neq i$. However, unlike previous approaches [17, 86] defined on a single modality (but similar to [80]), AVID uses multiple modalities, and thus can assume multiple forms as depicted in Figure 2.

1. **Self-AVID** requires instance discrimination within the same modality - $\mathbf{v}_i$ to $\bar{\mathbf{v}}_i$ and $\mathbf{a}_i$ to $\bar{\mathbf{a}}_i$. This is equivalent to prior work [17, 86] independently applied to the two modalities.
2. **Cross-AVID** optimizes for cross-modal discrimination, *i.e.*, the visual representation $\mathbf{v}_i$ is required to discriminate the accompanying audio memory $\bar{\mathbf{a}}_i$ and vice-versa.
3. **Joint-AVID** combines the Self-AVID and Cross-AVID objectives.

It is not immediately obvious what the relative advantages, if any, of these variants are. In §3.3, we provide an in-depth empirical study of the impact of these choices on the quality of the learned representations. We now describe the training procedure in detail.

3.2. AVID training procedure.

AVID is trained using a contrastive learning framework [27, 28], where instance representations are contrasted to those of other (negative) samples.

While various loss functions have been defined for contrastive learning [61, 73], we focus on noise contrastive estimation (NCE) [27]. Let $\bar{\mathbf{x}}_i$ denote the (memory) target representation for a sample s_i. The probability that a feature $\mathbf{x}$ belongs to sample s_i is modeled by a generalized softmax function

$$P(s_i|\mathbf{x}) = \frac{1}{NZ} \exp(\mathbf{x}^T \bar{\mathbf{x}}_i / \tau) \tag{1}$$

where $\bar{Z} = \frac{1}{N} \sum_{\bar{\mathbf{x}}} [\exp(\mathbf{x}^T \bar{\mathbf{x}} / \tau)]$ is the normalized partition function and τ is a temperature hyper-parameter that controls the softness of the distribution. In the case of AVID, $\mathbf{x}$ and $\bar{\mathbf{x}}$ may or may not be from the same modality.

The network f is trained to learn representations by solving multiple binary classification problems where it must choose its own target representation $\bar{\mathbf{x}}_i$ over representations $\bar{\mathbf{x}}_j$ in a negative set. The negative set consists of K 'other' instances drawn uniformly from $\mathcal{S}$, *i.e.*, $\mathcal{N}_i = \mathcal{U}(\mathcal{S})^K$. The probability of a feature $\mathbf{x}$ being from instance s_i as opposed to the instances from the uniformly sampled negative set $\mathcal{N}_i$ is given as $P(D = 1|\mathbf{x}, \bar{\mathbf{x}}_i) = \frac{P(s_i|\mathbf{x})}{P(s_i|\mathbf{x}) + K/N}$. The NCE loss is defined as the negative log-likelihood

$$\mathcal{L}_{\text{NCE}}(\mathbf{x}_i; \bar{\mathbf{x}}_i, \mathcal{N}_i) = -\log P(D = 1|\mathbf{x}_i, \bar{\mathbf{x}}_i) - \sum_{j \in \mathcal{N}_i} \log P(D = 0|\mathbf{x}_i, \bar{\mathbf{x}}_j), \tag{2}$$

where $P(D = 0|\cdot) = 1 - P(D = 1|\cdot)$.

The three variants of AVID depicted in Figure 2 are trained to optimize variations of the NCE loss of Equation 2,

Table 1: Variants of AVID. We observe that the Self-AVID and Joint-AVID variants that use within-modality instance discrimination perform poorly compared to Cross-AVID that uses only cross-modal instance discrimination.

Method	block1	block2	block3	block4	Best
Cross-AVID	**19.80**	**26.98**	**34.81**	**39.95**	**39.95**
Self-AVID	17.10	22.28	27.23	32.08	32.08
Joint-AVID	18.65	23.60	29.47	33.04	33.04

(a) Accuracy of linear probing on Kinetics.

	block1	block2	block3	block4	Best
Cross-AVID	**67.25**	**73.15**	**74.80**	**75.05**	**75.05**
Self-AVID	66.92	72.64	71.45	71.61	72.64
Joint-AVID	65.45	68.65	71.77	68.41	71.77

(b) Accuracy of linear probing on ESC.

by varying the target representations $\bar{\mathbf{x}}_i$.

$$\mathcal{L}_{\text{Self-AVID}} = \mathcal{L}_{\text{NCE}}(\mathbf{v}_i; \bar{\mathbf{v}}_i, \mathcal{N}_i) + \mathcal{L}_{\text{NCE}}(\mathbf{a}_i; \bar{\mathbf{a}}_i, \mathcal{N}_i) \quad (3)$$

$$\mathcal{L}_{\text{Cross-AVID}} = \mathcal{L}_{\text{NCE}}(\mathbf{v}_i; \bar{\mathbf{a}}_i, \mathcal{N}_i) + \mathcal{L}_{\text{NCE}}(\mathbf{a}_i; \bar{\mathbf{v}}_i, \mathcal{N}_i) \quad (4)$$

$$\mathcal{L}_{\text{Joint-AVID}} = \mathcal{L}_{\text{Self-AVID}}(\mathbf{v}_i, \mathbf{a}_i) + \mathcal{L}_{\text{Cross-AVID}}(\mathbf{v}_i, \mathbf{a}_i) \quad (5)$$

We analyze these variants next and show that the seemingly minor differences between them translate to significant differences in performance.

3.3. Analyzing AVID

We present experiments to analyze various properties of the AVID task and understand the key factors that enable the different variants of AVID to learn good representations.

Experimental Setup We briefly describe the experimental setup for analysis and provide the full details in the supplemental.

Pre-training Dataset. All models are trained using the Audioset dataset [24] which contains 1.8M videos focusing on audio events. We randomly subsample 100K videos from this dataset to train our models. We use input video and audio clips of 1 and 2-second duration, respectively. The video model is trained on 16 frames of size 112×112 with standard data augmentation [79]. We preprocess the audio by randomly sampling the audio within 0.5 seconds of the video and compute a log spectrogram of size 100×129 (100 time steps with 129 frequency bands).

Video and audio models. The video model is a smaller version of the R(2+1)D models proposed in [81] with 9 layers. The audio network is a 9 layer 2D ConvNet with batch normalization. In both cases, output activations are max-pooled, projected into a 128-dimensional feature using a multi-layer perceptron (MLP), and normalized into the unit sphere. The MLP is composed of three fully connected layers with 512 hidden units.

Pre-training details. AVID variants are trained to optimize the loss in Equations 3-5 with 1024 random negatives. In early experiments, we increased the number of negatives up to 8192 without seeing noticeable differences in performance. Following [86], we set the temperature hyper-parameter τ to 0.07, the EMA update constant to 0.5, and the normalized partition function $\bar{Z}$ is approximated during the first iteration and kept constant thereafter ($\bar{Z} = 2.2045$). All models are trained with the Adam optimizer [39] for 400 epochs with a learning rate of 1e-4, weight decay of 1e-5, and batch size of 256.

Downstream tasks. We evaluate both the visual and audio features using transfer learning.

- Visual Features: We use the Kinetics dataset [83] for action recognition. We evaluate the pre-trained features by linear probing [26, 91] where we keep the pre-trained network fixed and train linear classifiers. We report top-1 accuracy on held-out data by averaging predictions over 25 clips per video.
- Audio Features: We evaluate the audio features on the ESC-50 [66] dataset by training linear classifiers on fixed features from the pre-trained audio network. Similar to the video case, we report top-1 accuracy by averaging predictions over 25 clips per video.

Cross *vs*. within-modal instance discrimination

We study the three variants of AVID depicted in Figure 2 to understand the differences between cross-modal and within-modal instance discrimination and its impact on the learned representations. We evaluate the video and audio feature representations from these variants and report results in Table 1. We observe that Self-AVID is consistently outperformed by the Cross-AVID variant on both visual and audio tasks.

We believe the reason is that Self-AVID uses within-modality instance discrimination, which is an easier pretext task and can be partially solved by matching low-level statistics of the data [2, 16]. This hypothesis is supported by the fact that Joint-AVID, which combines the objectives of both Cross-AVID and Self-AVID, also gives worse performance than Cross-AVID. These results highlight that one *cannot* naively use within-modality instance discrimination when learning audio-visual representations. In contrast, Cross-AVID uses a "harder" cross-modal instance discrimination task where the video features are required to match the corresponding audio and vice-versa. As a result, it generalizes better to downstream tasks.

4. Beyond Instance Discrimination: Cross-Modal Agreement

We will show in §5 that Cross-AVID achieves state-of-the-art performance on action recognition downstream tasks. However, we identify three important limitations in the in-

stance discrimination framework of Equation 2 and the cross-modal loss of Equation 4.

1. **Limited to instances:** Instance discrimination does not account for interactions *between* instances. Thus, two semantically related instances are never grouped together and considered 'positives'.
2. **False negative sampling:** The negative set $\mathcal{N}_i$, which consists of *all* other instances s_j, may include instances semantically related to s_i. To make matters worse, contrastive learning requires a large number K of negatives, increasing the likelihood that semantically related samples are used as negatives. This contradicts the goal of representation learning, which is to generate similar embeddings of semantically related inputs.
3. **No within-modality calibration:** The Cross-AVID loss of Equation 4 does not directly optimize for visual similarity $\mathbf{v}_i^T\mathbf{v}_j$. In fact, as shown experimentally in §3.3, doing so can significantly hurt performance. Nevertheless, the lack of within-modality calibration is problematic, as good visual representations should reflect visual feature similarities.

4.1. Relating instances through agreements

We extend AVID with Cross-Modal Agreement (CMA) to address these shortcomings. CMA builds upon insights from prior work [70] in multi-view learning. We hypothesize that, if two samples are similar in *both* visual and audio feature space, then they are more likely to be semantically related than samples that agree in only one feature space (or do not agree at all). We thus consider instances that agree in both feature spaces to be 'positive' samples for learning representations. Similarly, examples with a poor agreement in either (or both) spaces are used as negatives. When compared to instance discrimination methods [17, 80, 86], CMA uses a larger positive set of semantically related instances and a more reliable negative set.

4.2. CMA Learning Objective

We define an agreement score for two instances s_i and s_j as

$$\rho_{ij} = \min(\mathbf{v}_i^T\mathbf{v}_j, \mathbf{a}_i^T\mathbf{a}_j). \tag{6}$$

This is large only when *both* the audio and video similarities are large. A set of positives and negatives is then defined per instance s_i. The positive set $\mathcal{P}_i$ contains the samples that are most similar to s_i in both spaces, while the negative set $\mathcal{N}_i$ is the complement of $\mathcal{P}_i$.

$$\mathcal{P}_i = \underset{j=1,\dots,N}{\text{TopK}}\,(\rho_{ij}) \qquad \mathcal{N}_i = \{j | s_j \in (\mathcal{S} \setminus \mathcal{P}_i)\} \tag{7}$$

Furthermore, CMA enables self-supervision beyond single instances. This is achieved with a generalization of the AVID task, which accounts for the correspondences of Equation 7. At training time, K_n negative instances are drawn per sample s_i from the associated negative set $\mathcal{N}_i$ to form set $\mathcal{N}_i' = \mathcal{U}(\mathcal{N}_i)^{K_n}$. The networks f_v, f_a are learned to optimize a combination of *cross-modal instance discrimination* and *within-modal positive discrimination* (wMPD). The former is encouraged through the Cross-AVID loss of Equation 4. The latter exploits the fact that CMA defines *multiple* positive instances $\mathcal{P}_i$, thus enabling the optimization of within-modality positive discrimination

$$\mathcal{L}_{\text{wMPD}} = \frac{1}{K_p}\sum_{p\in\mathcal{P}_i} \mathcal{L}_{\text{NCE}}(\mathbf{v}_i; \bar{\mathbf{v}}_p, \mathcal{N}_i') + \mathcal{L}_{\text{NCE}}(\mathbf{a}_i; \bar{\mathbf{a}}_p, \mathcal{N}_i'). \tag{8}$$

Note that, unlike the Self-AVID objective of Equation 3, this term calibrates within-modal similarities between positive samples. This avoids within-modal comparisons to the instance itself, which was experimentally shown to produce weak representations in §3.3. We then minimize the weighted sum of the two losses

$$\mathcal{L}_{\text{CMA}} = \mathcal{L}_{\text{Cross-AVID}}(\mathbf{v}_i, \mathbf{a}_i) + \lambda\mathcal{L}_{\text{wMPD}}(\mathbf{v}_i, \mathbf{a}_i), \tag{9}$$

where $\lambda > 0$ is an hyper-parameter that controls the weight of the two losses.

Implementation. After Cross-AVID pre-training, cross-modal disagreements are corrected by finetuning the audio and video networks to minimize the loss in Equation 9. Models are initialized with the Cross-AVID model at epoch 200, and trained for 200 additional epochs. We compare these models to a Cross-AVID model trained for 400 epochs, thus controlling for the total number of parameter updates. For each sample, we find 32 positive instances using the CMA criterion of Equation 7 applied to video and audio memory bank representations. For efficiency purposes, the positive set is updated every 50 epochs. In each iteration, 1024 negative memories (not overlapping with positives) were sampled. These positive and negative memories were then used to minimize the CMA loss of Equations 8-9. For evaluation purposes, we use the same protocol as in §3.3.

4.3. Analyzing CMA

The CMA objective consists of two terms that optimize cross-modal (Equation 4) and within-modal (Equation 8) similarity. We observed in §3.3 that within-modal comparisons for instance discrimination result in poor visual representations due to the relatively easy task of self-discrimination. Intuitively, since CMA identifies groups of instances ($\mathcal{P}_i$) that are likely related, calibrating within-modal similarity within these groups (instead of within the instance itself) should result in a better visual representation. To study this, we use CMA to obtain a positive set $\mathcal{P}_i$ and analyse the CMA objective of Equation 9 by evaluating with different values of the hyper-parameter λ. The results shown in Figure 3 validates the advantages of CMA over Cross-AVID.

CMA calibration. To understand the effect of the CMA procedure on within-modal similarities, we analyzed the embedding space defined by memory bank representations obtained with AVID and CMA trained on the Kinetics dataset. Since representations are restricted to the unit sphere (due to normalization), the average inner-product between two randomly chosen samples should be 0 (assuming a uniform distribution of samples over the sphere). However, when training with Cross-AVID, the average inner-product is 0.23. This means that Cross-AVID learns collapsed representations (*i.e.* features are on average closer to other random features than the space permits). This is likely due to the lack of within-modal negatives when training for cross-modal discrimination. By seeking within modal-discrimination of positive samples, CMA effectively addresses the feature collapsing problem observed for Cross-AVID, and yields an average dot-product between random memories of 0 as expected.

CMA *vs*. within-modal expansion. CMA expands the positive set $\mathcal{P}_i$ to include instances that *agree in both* video and audio spaces. We inspected whether modeling this agreement is necessary for relating instances by exploring alternatives that do not model agreements in both spaces (see Figure 4a). We consider alternatives that expand the set $\mathcal{P}_i$ by looking at instances that are similar in 1) only the audio space; 2) only the video space; or 3) either video or audio space. Each method in Figure 4a is trained to optimize the objective of Equation 9 with the corresponding $\mathcal{P}_i$. We also compare against the Cross-AVID baseline that uses only the instance itself as the positive set. Transfer performance is reported in Figure 4b.

Compared to Cross-AVID, expanding the set of positives using only audio similarity (third row) hurts performance on Kinetics, and relying on video similarities alone (second row) only provides marginal improvements. We believe that expanding the set of positives only based on visual similarity does not improve the performance of visual features since the positives are *already close* in the feature space, and do not add extra information. CMA provides consistent gains over all methods on Kinetics, suggesting that modeling *agreement* can provide better positive sets for representation learning of visual features.

Qualitative Understanding. We show examples of positive and negative samples found by CMA in Figure 5 and observe that CMA can group together semantically related concepts. As it uses agreement between both spaces, visually similar concepts, like 'ambulance' and 'bus' (second row), can be distinguished based on audio similarity. This leads to more precise positive sets $\mathcal{P}_i$, as can be verified by inspecting the precision@K of $\mathcal{P}_i$ measured against ground truth labels (Figure 4c). CMA consistently finds more precise positives compared to within-modal expansion methods showing the advantages of modeling agreement.

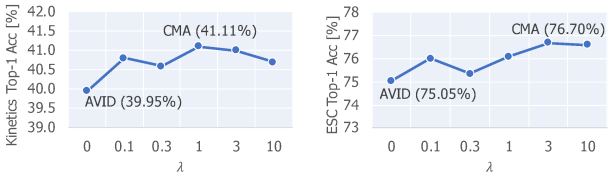


Figure 3: Ablation of CMA objective. Impact of within-modal positive sample discrimination. A network is pre-trained for different values of hyper-parameter λ in Equation 9, and then evaluated by linear probing on the Kinetics and ESC datasets. Positive sample discrimination can further improve the performance of Cross-AVID.

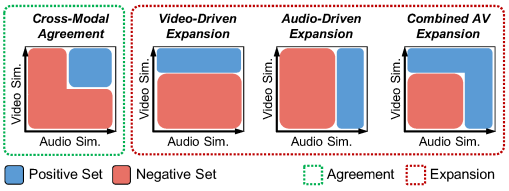
5. Cross-AVID and CMA at scale

Previous sections provide experimental validation for the proposed Cross-AVID and CMA procedures when training on a medium-sized dataset (100K videos from Audioset). We now study the proposed methods on large-scale datasets. We also compare Cross-AVID and CMA to prior work, including video-based self-supervised learning methods [29, 49, 54, 87], and methods that leverage the natural correspondence between audio and video [2, 41, 62].

Experimental setup. We briefly describe the experimental setup, and refer the reader to supplementary material for full details. We use the 18-layer R(2+1)D network of [81] as the video encoder and a 9-layer (2D) CNN with batch normalization as the audio encoder. Models are trained on Kinetics-400 [83] and the full Audioset [24] datasets, containing 240K and 1.8M video instances, respectively. Video clips composed of 8 frames of size 224×224 are extracted at a frame rate of 16fps with standard data augmentation procedures [79]. Two seconds of audio is randomly sampled within 0.5 seconds of the video at a 24kHz sampling rate, and spectrograms of size 200×257 (200 time steps with 257 frequency bands) are used as the input to the audio network. For Cross-AVID, the cross-modal discrimination loss of Equation 4 is optimized with $K = 1024$ negative instances. We then find 128 positive instances for each sample using cross-modal agreements (Equation 7), and optimize the CMA criterion of Equation 9 with $K_p = 32$ positives, $K_n = 1024$ negatives and $\lambda = 1.0$. Video representations are evaluated on action recognition (§5.1), and audio representations on sound classification (§5.2).

5.1. Action recognition

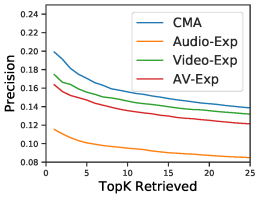
We first evaluate the visual representations learned by Cross-AVID and AVID+CMA by training a linear classifier for the task of action recognition on the Kinetics dataset. The top-1 accuracy is reported for clip and video-level pre-



(a) Positive and negative sets of 'agreement' and 'expansion' methods.

Method	block1	block2	block3	block4	Best
Cross-AVID (Base)	19.80	26.98	34.81	39.95	39.95
Base + Video-Exp.	19.93	27.39	35.64	40.17	40.17
Base + Audio-Exp.	20.14	27.28	35.68	39.62	39.62
Base + AV Exp	20.04	27.61	36.14	40.58	40.58
Base + CMA	**20.16**	**27.98**	**36.98**	**41.11**	**41.11**

(b) Top-1 accuracy of linear probing on Kinetics.



(c) Precision@K.

Figure 4: Cross-Modal Agreement *vs*. Within-modality Expansion We study the importance of modeling *agreement* between video and audio similarities. We compare CMA to expansion methods that relate instances without modeling agreement (4a). CMA enables better transfer for action recognition (4b). Expansion methods generate agreements of worse precision (4c).

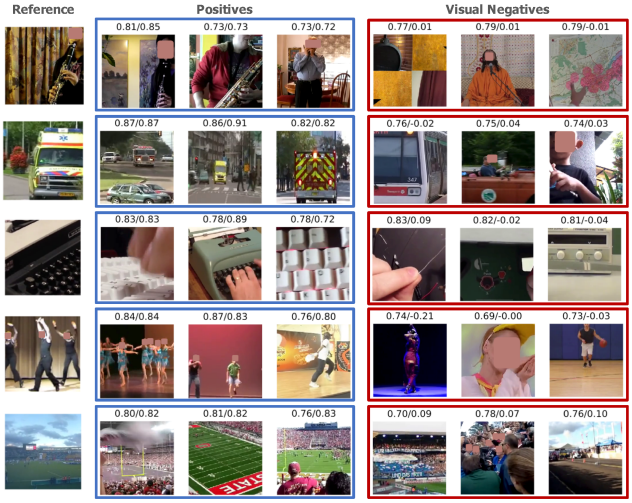


Figure 5: Examples extracted by the CMA procedure. For each reference image, we show three images in their positive sets (Equation 7). We also show three negatives that were rejected from the positive set due to low audio similarity. Each image is annotated with the video/audio similarity to the reference.

Pretraining DB	Kinetics		Audioset	
Method \ Metric	Clip@1	Video@1	Clip@1	Video@1
Cross-AVID	33.3	43.1	35.2	46.6
AVID+CMA	**35.1**	**44.5**	**37.4**	**48.9**

Table 2: Top-1 accuracy of linear probing on Kinetics.

dictions. Clip-level predictions are obtained from a single 8-frame clip, while video-level predictions are computed by averaging clip-level predictions from 10 clips uniformly sampled from the whole video. The results shown in Table 2 clearly demonstrate the advantage of calibrating AVID representations using the CMA procedure, yielding significant gains across both metrics and pretraining datasets. These results demonstrate the value of the CMA procedure in large-scale datasets, thus showing that its effect goes beyond a simple regularization procedure to prevent overfitting.

To compare to prior work, we follow [29, 41, 80] and evaluate visual representations on the UCF-101 [76] and HMDB-51 [42] datasets, by full network fine-tuning. Due to the large variability of experimental setups used in the literature, it is unrealistic to provide a direct comparison to all methods, as these often use different network encoders trained on different datasets with input clips of different lengths. To increase the range of meaningful comparisons, we fine-tuned our models using clips with both 8 and 32 frames. At inference time, video-level predictions are provided by averaging clip-level predictions for 10 uniformly sampled clips [41]. We report top-1 accuracy averaged over the three train/test splits provided with the original datasets.

Table 3 compares the transfer performance of Cross-AVID and CMA with previous self-supervised approaches. To enable well-grounded comparisons, we also list for each method the pre-training dataset and clip dimensions used while finetuning on UCF and HMDB. Despite its simplicity, Cross-AVID achieves state-of-the-art performance for equivalent data settings in most cases. In particular, when pre-trained on Audioset, Cross-AVID outperformed other audio-visual SSL methods such as L3 and AVTS by at least 1.0% on UCF and 2.5% on HMDB. Similar to Cross-AVID, L3 and AVTS propose to learn audio-visual representations by predicting whether audio/video pairs are in-sync. However, these methods optimize for the audiovisual correspondence task, which fails to reason about the data distribution at large. Cross-AVID also outperformed the concurrently proposed XDC [1] under equivalent data settings. When pre-trained on Audioset and finetuned on UCF with 32 frames, XDC [1] does report higher accuracy, but the model was pretrained and finetuned using 32 frames, while we pretrain using only 8 frames. It should be noted that, when pre-training and finetuning with clips of 8 frames, Cross-AVID outperforms XDC by 3.4% (84.9% vs 88.3%). CMA further improves the performance of Cross-AVID on all settings considered (*i.e.*, using both Kinetics and Audioset pretraining datasets, and evaluating on UCF and HMDB). We observed, however, that the improvements of CMA over Cross-AVID are smaller under the fine-tuning protocol than the linear evaluation of Table 2. Prior work [26, 91] observes that full fine-tuning significantly modifies the visual features and tests

Method	Pretraining DB	Finetune Input Size	UCF	HMDB
Shuffle&Learn [54]	UCF	1×227^2	50.2	18.1
OPN [49]	UCF	1×227^2	56.3	23.8
ST Order [9]	UCF	1×227^2	58.6	25.0
CMC [80]	UCF	1×227^2	59.1	26.7
3D-RotNet [37]	Kinetics400	16×112^2	62.9	33.7
ClipOrder [87]	Kinetics400	16×112^2	72.4	30.9
DPC [29]	Kinetics400	25×128^2	75.7	35.7
CBT [78]	Kinetics400	16×112^2	79.5	44.6
L3* [2]	Kinetics400	16×224^2	74.4	47.8
AVTS [41]	Kinetics400	25×224^2	85.8	56.9
XDC [1]	Kinetics400	8×224^2	74.2	39.0
	Kinetics400	32×224^2	86.8†	52.6†
Cross-AVID (ours)	Kinetics400	8×224^2	82.3	49.1
	Kinetics400	32×224^2	86.9	59.9
AVID+CMA (ours)	Kinetics400	8×224^2	83.7	49.5
	Kinetics400	32×224^2	**87.5**	**60.8**
L3* [2]	Audioset	16×224^2	82.3	51.6
Multisensory [62]	Audioset	64×224^2	82.1	–
AVTS [41]	Audioset	25×224^2	89.0	61.6
XDC [1]	Audioset	8×224^2	84.9	48.8
	Audioset	32×224^2	93.0†	63.7†
Cross-AVID (ours)	Audioset	8×224^2	88.3	57.5
	Audioset	32×224^2	91.0	64.1
AVID+CMA (ours)	Audioset	8×224^2	88.6	57.6
	Audioset	32×224^2	**91.5**	**64.7**

Table 3: Top-1 accuracy on UCF and HMDB by full network finetuning with various pre-training datasets and clips of different sizes. Methods were organized by pre-training dataset. The method with the best performance is indicated in bold face, and second best is underlined. *Re-implemented by us. †Obtained by pre-training and finetuning with larger 32×224^2 inputs (we only pre-train on 8×224^2 inputs).

the network initialization aspect of pre-training rather than the semantic quality of the representation. Thus, we believe that the feature calibration benefits of CMA are diminished under the full finetuning protocol.

5.2. Sound recognition

Audio representations are evaluated on the ESC-50 [66] and DCASE [77] datasets by linear probing [26] for the task of sound recognition. Following [41], both ESC and DCASE results are obtained by training a linear one-vs-all SVM classifier on the audio representations generated by the pre-trained models at the final layer before pooling. For training, we extract 10 clips per sample on the ESC dataset and 60 clips per sample on DCASE [41]. At test time, sample level predictions are obtained by averaging 10 clip level predictions, and the top-1 accuracy is reported in Table 4. For the ESC dataset, performance is the average over the 5 original train/test splits. Similarly to video, audio representations learned by Cross-AVID and CMA outperform prior work, outperforming ConvRBM on the ESC dataset by 2.7% and AVTS on DCASE by 3%.

Method	Pretraining DB	ESC	DCASE
RandomForest [66]	None	44.3	–
ConvNet [65]	None	64.5	–
ConvRBM [71]	None	86.5	–
SoundNet [4]	Flickr-SoundNet	74.2	88
L3 [2]	Flickr-SoundNet	79.3	93
AVTS [41]	Kinetics	76.7	91
XDC [1]	Kinetics	78.5	–
Cross-AVID (Ours)	Kinetics	77.6	**93**
AVID+CMA (Ours)	Kinetics	**79.1**	**93**
AVTS [41]	Audioset	80.6	93
XDC [1]	Audioset	85.8	–
Cross-AVID (Ours)	Audioset	**89.2**	**96**
AVID+CMA (Ours)	Audioset	89.1	**96**

Table 4: Top-1 accuracy of linear classification on ESC-50 and DCASE datasets. Methods are organized by pre-training dataset. The method with the best performance is indicated in bold face, and second best is underlined.

6. Discussion

We proposed a self-supervised method to learn visual and audio representations by contrasting visual representations against multiple audios, and vice versa. Our method, Audio-Visual Instance Discrimination (AVID) builds upon recent advances in contrastive learning [80, 86] to learn state-of-the-art representations that outperform prior work on action recognition and sound classification. We propose and analyze multiple variants of the AVID task to show that optimizing for cross-modal similarity and not within-modal similarity matters for learning from video and audio.

We also identified key limitations of the instance discrimination framework and proposed CMA to use agreement in the video and audio feature spaces to group together related videos. CMA helps us relate *multiple* instances by identifying more related videos. CMA also helps us reject 'false positives', *i.e.*, videos that are similar visually but differ in the audio space. We show that using these groups of related videos allows us to optimize for within-modal similarity, in addition to cross-modal similarity, and improve visual and audio representations. The generalization of CMA suggests that cross-modal agreements provide non-trivial correspondences between samples and are a useful way to learn improved representations in a multi-modal setting.

Acknowledgements

We are grateful to Rob Fergus and Laurens van der Maaten for their feedback and support; Rohit Girdhar for feedback on the manuscript; and Bruno Korbar for help with the baselines.

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 2020. 2, 7, 8

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 4, 6, 8

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2, 8

[5] Francis R Bach, Gert RG Lanckriet, and Michael I Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceeding of the International Conference on Machine Learning (ICML)*, 2004. 2

[6] Steffen Bickel and Tobias Scheffer. Multi-view clustering. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2004. 2

[7] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Annual Conference on Computational Learning Theory*, 1998. 2

[8] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *Proceeding of the International Conference on Machine Learning (ICML)*, 2017. 1

[9] Uta Buchler, Biagio Brattoli, and Bjorn Ommer. Improving spatiotemporal self-supervision by deep reinforcement learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 8

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[11] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 1

[12] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2016. 2

[13] Virginia R de Sa. Learning classification with unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1994. 1, 2

[14] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015. 1

[15] Tom Diethe, David R Hardoon, and John Shawe-Taylor. Multiview fisher discriminant analysis. In *Workshop in Advances in Neural Information Processing Systems (NeurIPS)*, 2008. 2

[16] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015. 1, 4

[17] Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(9):1734–1747, 2016. 1, 2, 3, 5

[18] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[19] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning by rotation feature decoupling. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[20] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[21] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 2

[22] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[23] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *IEEE International Conference on Computer Vision*, pages 3879–3888, 2019. 2

[24] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 4, 6

[25] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 1

[26] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 4, 7, 8

[27] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *ICAIS*, 2010. 3

[28] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. 1, 2, 3

[29] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Workshop on Large Scale Holistic Video Understanding, ICCV*, 2019. 1, 2, 6, 7, 8

[30] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[31] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[32] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*. PMLR, 2020. 2

[33] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018. 2

[34] Chih-Hui Ho and Nuno Vasconcelos. Contrastive learning with adversarial examples. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[35] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information distillation for unsupervised image segmentation and clustering. *arXiv preprint arXiv:1807.06653*, 2018. 1

[36] Huaizu Jiang, Gustav Larsson, Michael Maire Greg Shakhnarovich, and Erik Learned-Miller. Self-supervised relative depth learning for urban scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[37] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*, 2018. 1, 8

[38] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 2

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[40] Marius Kloft and Gilles Blanchard. The local rademacher complexity of lp-norm multiple kernel learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011. 2

[41] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 2, 6, 7, 8

[42] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *2011 International Conference on Computer Vision (ICCV)*. IEEE, 2011. 1, 7

[43] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011. 2

[44] Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5(Jan):27–72, 2004. 2

[45] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[46] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[47] Quoc V Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S Corrado, Jeff Dean, and Andrew Y Ng. Building high-level features using large scale unsupervised learning. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013. 1

[48] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007. 1

[49] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 6, 8

[50] Fan Ma, Deyu Meng, Qi Xie, Zina Li, and Xuanyi Dong. Self-paced co-training. In *Proceeding of the International Conference on Machine Learning (ICML)*, 2017. 2

[51] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-svms for object detection and beyond. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 3

[52] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *ICANN*. Springer, 2011. 1

[53] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[54] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1, 6, 8

[55] Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In *Proceeding of the International Conference on Machine Learning (ICML)*, 2009. 1

[56] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[57] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[58] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. 1

[59] Bruno A Olshausen. Sparse coding of time-varying natural images. In *Proc. of the Int. Conf. on Independent Component Analysis and Blind Source Separation*, 2000. 1

[60] Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607, 1996. 1

[61] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2, 3

[62] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision*

(ECCV), 2018. 1, 2, 6, 8

[63] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[64] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[65] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015. 8

[66] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the ACM International Conference on Multimedia*, 2015. 4, 8

[67] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[68] Novi Quadrianto and Christoph Lampert. Learning multi-view neighborhood preserving projections. In *Proceeding of the International Conference on Machine Learning (ICML)*, 2011. 2

[69] Marc'aurelio Ranzato, Fu Jie Huang, Y-Lan Boureau, and Yann LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 1

[70] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *WACV*, 2005. 5

[71] Hardik B Sailor, Dharmesh M Agrawal, and Hemant A Patil. Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification. In *InterSpeech*, 2017. 8

[72] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455, 2009. 1

[73] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3

[74] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[75] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2018. 2

[76] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. Technical Report CRCV-TR-12-01, University of Central Florida, 2012. 1, 7

[77] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015. 8

[78] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019. 8

[79] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 4, 6

[80] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Workshop on Self-Supervised Learning, ICML*, 2019. 1, 2, 3, 5, 7, 8

[81] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4, 6

[82] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceeding of the International Conference on Machine Learning (ICML)*. ACM, 2008. 1

[83] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *arXiv:1705.06950*, 2017. 1, 4, 6

[84] Wei Wang and Zhi-Hua Zhou. Analyzing co-training style algorithms. In *Proceeding of the European Conference on Machine Learning (ECML)*. Springer, 2007. 2

[85] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015. 1

[86] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 4, 5, 8

[87] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6, 8

[88] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[89] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[90] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1

[91] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 4, 7

[92] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Tor-

ralba. The sound of motions. In *IEEE International Conference on Computer Vision*, 2019. 2

[93] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[94] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 2