Robust Audio-Visual Instance Discrimination

Pedro Morgado* UC San Diego

Ishan Misra Facebook AI Research

Nuno Vasconcelos UC San Diego

Abstract

We present a self-supervised learning method to learn audio and video representations. Prior work uses the natural correspondence between audio and video to define a standard cross-modal instance discrimination task, where a model is trained to match representations from the two modalities. However, the standard approach introduces two sources of training noise. First, audio-visual correspondences often produce faulty positives since the audio and video signals can be uninformative of each other. To limit the detrimental impact of faulty positives, we optimize a weighted contrastive learning loss, which down-weighs their contribution to the overall loss. Second, since self-supervised contrastive learning relies on random sampling of negative instances, instances that are semantically similar to the base instance can be used as faulty negatives. To alleviate the impact of faulty negatives, we propose to optimize an instance discrimination loss with a soft target distribution that estimates relationships between instances. We validate our contributions through extensive experiments on action recognition tasks and show that they address the problems of audio-visual instance discrimination and improve transfer learning performance.

1. Introduction

Self-supervised representation learning aims to learn feature representations that can transfer to downstream tasks without costly human annotations. Many recent self-supervised methods [11, 36, 51, 14, 78, 74] use a variant of the instance discrimination framework [81, 22], which matches features from multiple views/augmentations of the *same* instance, while distinguishing these features from those of other instances. This often relies on a contrastive loss [31], where different augmentations are considered 'positives' and other samples 'negatives.'

Cross-modal instance discrimination (xID) extends instance discrimination to the realm of multiple modalities, where data modalities, such as video, audio, or text, act as the different 'views' of an instance. Since there is a strong correlation between audio and visual events (*e.g.*, the sound of an instrument or a baseball match), audio-visual instance discrimination has gained popularity [5, 58, 41, 56, 64, 3, 61, 2]. Representations learned by these methods show promising performance on tasks like action recognition and environmental sound classification. xID methods rely on two key assumptions - (1) the audio and video of a sample are informative of each other, *i.e.*, positives; (2) the audio and video of all other samples are not related, *i.e.*, negatives. In practice, both these assumptions are too strong and do not hold for a significant amount of real-world data. This results in *faulty positive* samples that are not related to each other and *faulty negative* samples that are semantically related.

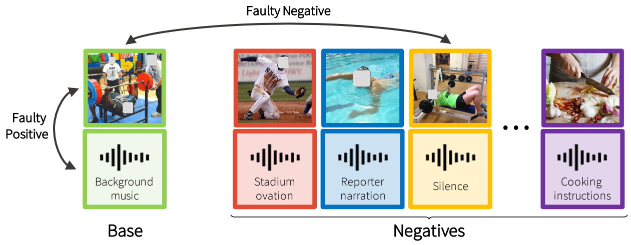

Figure 1: Example of a positive audio/video pair and negative instances used for contrastive learning. Audio-visual signals may not semantically correspond to each other, such as the pairing weightlifting video/background music shown in green, which leads to faulty positives. Due to random sampling, semantically similar instances can also appear as faulty negatives, *e.g.* a second weightlifting video in yellow. Faulty positive and negative samples are a common occurrence in audio-visual contrastive learning and hurt representation learning.

Figure 1 shows examples of these faulty correspondences. Videos where the audio is uninformative of the visual content can lead to faulty positives, *e.g.*, videos containing audio from sources outside of the camera field-of-view or containing post-edited sounds like a soundtrack. Similarly, random negative sampling can produce faulty negatives, *i.e.*, negative samples that are semantically related to the positive. These faulty correspondences undermine the primary goal of representation learning, *i.e.*, to ensure that similar instances have similar feature representations. As we show empirically in Figure 7 and Table 1, they can hurt representation learning and degrade downstream performance. Thus, we believe cross-modal learning should be seen as a problem

[*] Contacting author.

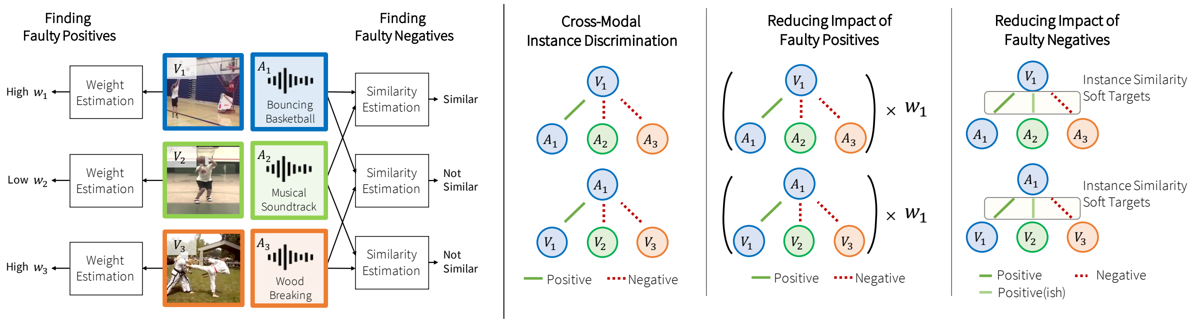

Figure 2: Comparison between standard cross-modal instance discrimination (xID) and the proposed procedure. In xID, samples contribute equally to the loss, and all instances other than themselves are treated as negatives. The proposed method addresses the two main sources of noisy training signals: faulty positives and faulty negatives. Faulty positives are discounted by down-weighting instances with poor audio-visual correspondence. Faulty negatives are addressed by optimizing the loss over a soft target distribution that encodes instance similarity.

of learning with *noisy targets*. This raises the question of how to identify faulty positive and negative samples in the absence of human annotations.

We propose to use cross-modal information during self-supervised training to detect both faulty positive and negative instances. This is done by estimating the quality of the audio-visual correspondence of each instance and optimizing a weighted contrastive learning loss that down-weighs the contribution of faulty positive examples. To address faulty negatives, we estimate the similarity *across* instances to compute a soft target distribution over instances. The model is then tasked to match this distribution. As a result, instances with enough evidence of similarity are no longer used as negatives and may even be used as positives.

The contributions of this work are as follows (Figure 2). We identify two sources of training noise in cross-modal learning: instances with weak cross-modal correspondence, which create *faulty positives*, and the sampling of semantically similar instances as negatives, which create *faulty negatives*. We show that removing faulty positives and negatives using an oracle can significantly improve the performance of a state-of-the-art xID method [56]. We then propose a mechanism to replace the oracle and a robust cross-modal instance discrimination loss that limits the impact of faulty correspondences. The effectiveness of the proposed method is demonstrated on several downstream tasks.

2. Related work

Self-supervised representation learning aims to learn representations by solving pretext tasks defined from the data alone, *i.e.* without human annotations. In computer vision, pretext tasks involve reasoning about spatial context [57, 21, 39, 60, 30, 34, 65], temporal context [52, 45, 80, 39, 53, 33, 23, 77, 9, 34, 35, 65], other visual properties such as hue, brightness and flow [20, 43, 86, 44, 87, 74, 64], or clusters of features [10, 7, 11, 78]. One promising technique is the instance discrimination task proposed in [81, 22] and further explored in [36, 51, 14, 78, 82]. However, contrastive learning from a single modality requires heavy data augmentations to generate distinct views. Instead, we focus on cross-modal instance discrimination, which avoids this issue by generating views from different modalities.

Representation learning from audio-visual correspondences: Since, in video, the audio is naturally paired and synced with the visual component, audio-visual correspondences have been used to draw direct supervision for several tasks, such as visually guided-source separation and localization [26, 28, 91, 90, 25, 68], visually guided audio spatialization [55, 27], audio-visual embodied navigation [13], lip-speech synchronization [18] and audio-visual speech recognition [1, 17].

In the context of contrastive learning, audio-visual correspondences are used to generate alternative views of an instance. While this has been known for a long time [19], self-supervised audio-visual representation learning gained popularity in recent years. For example, [5, 4] propose to learn representations by solving a *binary* classification problem that identifies audio and video clips belonging to the same instance. [41, 58] predict if audio and video clips are temporally synchronized, and [54] predicts if audio and video clips extracted from a 360 video are spatially aligned. [56, 61] improve upon the audio-visual correspondence problem [5] by posing it as a cross-modal instance discrimination task, where instances are contrasted to a large number of negatives. As a result, [56, 61] achieve impressive performance on downstream tasks such as action recognition.

In this work, we address two issues inherent to cross-modal instance discrimination, namely the detrimental impact of faulty positives and negatives. Recently, [3, 8] proposed to learn representations by iteratively clustering the audio and visual representations and seeking to predict cluster assignments from the opposite modality. While clustering can also discourage faulty negatives from acting as repelling forces, our method accomplishes this by optimizing a simple instance discrimination loss with soft targets, thus avoiding the significant computational overhead of clustering.

Supervised learning from noisy labels. Our work is closely related to supervised learning from noisy labels [66, 88, 62, 32, 47]. Since label collection is expensive and time-consuming, scaling human annotation to large datasets often requires the use of non-experts or non-curated labels such as user tags, which are prone to noise. Since deep neural networks can easily overfit to noisy labels [84], this results in poor generalization. Several techniques have been developed to increase the robustness of learning algorithms to label noise, including losses that reduce the impact of outliers [29, 88, 79], loss correction approaches that model the sources of label noise [62, 37, 12, 66, 6, 49, 70], meta-learning procedures that learn how to correct the sources of label noise [47, 67, 46, 69, 89] and regularization procedures tailored to lower the impact of noise [85, 63]. We refer the reader to [71, 24] for a detailed survey of prior work on learning with label noise. In this work, we show that cross-modal instance discrimination should be seen as a problem of learning with noisy targets. However, instead of the class mislabeling, we identify two main sources of noise for cross-modal instance discrimination (faulty positives and faulty negatives) and propose an algorithm to mitigate them.

3. Analysis: Instance Discrimination

We analyze the cross-modal instance discrimination method [56, 74, 61] and show that faulty positives and negatives have a disproportionately large contribution to the training updates. Additionally, in Table 1, we document the detrimental empirical effects of faulty samples.

Cross-Modal Instance Discrimination Consider a dataset $\mathcal{D} = \{(v_i, a_i)_{i=1}^N\}$ containing N samples (or instances) of video v_i and audio a_i. Cross-modal instance discrimination uses a contrastive loss [31] to learn video and audio encoders, $f_v(\cdot)$ and $f_a(\cdot)$, so as to align the two modalities belonging to the same instance [74, 56, 61] by minimizing

$$L_{\text{xID}}(\mathbf{v}_i, \mathbf{a}_i) = -\log P(\bar{\mathbf{a}}_i|\mathbf{v}_i;\tau) - \log P(\bar{\mathbf{v}}_i|\mathbf{a}_i;\tau) \tag{1}$$

$$\text{where} \quad P(\bar{\mathbf{t}}_i|\mathbf{s}_i;\tau) = \frac{\exp(\mathbf{s}_i^T\bar{\mathbf{t}}_i/\tau)}{\sum_k \exp(\mathbf{s}_i^T\bar{\mathbf{t}}_k/\tau)}, \tag{2}$$

where $\mathbf{v}_i = f_v(v_i)$ and $\mathbf{a}_i = f_a(a_i)$ are visual and audio features normalized to the unit sphere, $\bar{\mathbf{v}}_i$ and $\bar{\mathbf{a}}_i$ are target representations, and τ is a temperature hyper-parameter. Prior works differ by the type of target representations employed. For example, $\bar{\mathbf{v}}_i$ and $\bar{\mathbf{a}}_i$ can be entries of a memory bank as in [56, 81], the network representations themselves $\bar{\mathbf{v}}_i = f_v(v_i)$ and $\bar{\mathbf{a}}_i = f_a(a_i)$ as in SimCLR [14], the outputs of momentum encoders as in MoCo [36], or the centroids of an online clustering procedure as in SwAV or CLD [11, 78]. In this work, we build on the Audio-Visual Instance Discrimination (AVID) method of [56], focusing on target representations sampled from a memory bank. However, the principles introduced below can also be applied to SimCLR, MoCo or SwAV style targets.

Faulty positives and negatives in practice. The contrastive loss of Equation 1 is minimized when audio and visual representations from the same instance are aligned (dot-product similarities $\mathbf{v}_i^T\bar{\mathbf{a}}_i$ and $\mathbf{a}_i^T\bar{\mathbf{v}}_i$ as close to 1 as possible), and representations from different instances are far apart. In practice, however, the two modalities are not informative of each other for a significant number of instances (see Figure 1). We refer to these unclear correspondences as *faulty positives*.[1] On the other hand, a significant number of contrastive learning negatives are semantically similar to the base instance. We term these semantically similar negatives as *faulty negatives* since they should ideally be used as positives.

Figure 3 shows the histogram of similarities $\bar{\mathbf{v}}_i^T\bar{\mathbf{a}}_i$ after training an audio-visual model with the loss of Equation 1. As can be seen, instances with higher scores tend to have stronger correspondences (*i.e.* the audio and video signals are informative of each other). Instances where the two modalities are uninformative of each other tend to have lower scores and are generally faulty positives. On the other hand, Figure 4 shows the histograms of similarities between a video i and negatives j. As can be seen, faulty negatives tend to occur for negatives j with high similarity $\bar{\mathbf{v}}_i^T\bar{\mathbf{a}}_j$.

How do faulty positives and negatives affect learning? Faulty positives and negatives have a *disproportionately large* contribution to the training updates. To see this, examine the gradients that are computed when optimizing Equation 1. The partial derivatives are given as

$$-\frac{\partial L_{\text{xID}}}{\partial \mathbf{v}_i} = \underbrace{\frac{\bar{\mathbf{a}}_i}{\tau}(1-P(\bar{\mathbf{a}}_i|\mathbf{v}_i))}_{\text{Attraction force}} - \underbrace{\sum_{n\neq i}\frac{\bar{\mathbf{a}}_n}{\tau}P(\bar{\mathbf{a}}_n|\mathbf{v}_i)}_{\text{Repulsion force}} \tag{3}$$

$$-\frac{\partial L_{\text{xID}}}{\partial \mathbf{a}_i} = \underbrace{\frac{\bar{\mathbf{v}}_i}{\tau}(1-P(\bar{\mathbf{v}}_i|\mathbf{a}_i))}_{\text{Attraction force}} - \underbrace{\sum_{n\neq i}\frac{\bar{\mathbf{v}}_n}{\tau}P(\bar{\mathbf{v}}_n|\mathbf{a}_i)}_{\text{Repulsion force}}. \tag{4}$$

Intuitively, the target representations $\bar{\mathbf{v}}_i$ and $\bar{\mathbf{a}}_i$ of the instance itself act as 'attraction points' for the encoder of the opposing modality, while the target representations of other (negative) instances, $\bar{\mathbf{v}}_n$ and $\bar{\mathbf{a}}_n$, act as 'repelling points'. For example, in Equation 3, the negative gradient pushes $\mathbf{v}_i$ toward $\bar{\mathbf{a}}_i$ and away from $\bar{\mathbf{a}}_n, n \neq i$. The attraction forces are weighed by the complement of the prediction confidence, *i.e.*, $1 - P(\bar{\mathbf{v}}_i|\mathbf{a}_i)$ or $1 - P(\bar{\mathbf{a}}_i|\mathbf{v}_i)$. When positive samples are faulty, these gradients lead to noisy training signals. As show in Figure 3, faulty positives tend to have lower similarities and thus less confident predictions. As a result, the cross-modal loss of Equation 1 assigns stronger gradients to faulty positive samples. On the other hand, the repelling forces of negative instances are also weighted by the likelihood of matching the base sample, *i.e.* $P(\bar{\mathbf{v}}_n|\mathbf{a}_i)$

[1] We prefer 'faulty positives' over 'false positives' to distinguish from supervised learning where one has access to labels.

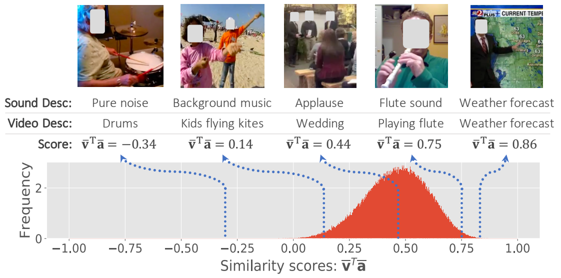

Figure 3: Faulty positives in a pretrained cross-modal model. Histogram of similarity scores $\bar{\mathbf{v}}_i^T\bar{\mathbf{a}}_i$ between video and audio representations, and examples obtained at various points of the distribution. We describe both the sound and video content in the plot. Examples with lower similarity scores contain audio that is less predictive of the video content, which creates faulty positives for training.

and $P(\bar{\mathbf{a}}_n|\mathbf{v}_i)$. However, as shown in Figure 4, faulty negatives tend to have high similarity scores, leading to high posteriors $P(\bar{\mathbf{v}}_n|\mathbf{a}_i)$ and $P(\bar{\mathbf{a}}_n|\mathbf{v}_i)$. Thus, the targets $\bar{\mathbf{v}}_n$ and $\bar{\mathbf{a}}_n$ of faulty negatives act as *strong* repelling forces for $\mathbf{a}$ and $\mathbf{v}$ (see Equation 3-4), even though they should ideally be close in feature space.

4. Robust audio-visual representation learning

We have seen that contrastive learning places too much emphasis on the impossible goals of bringing together the audio-visual components of faulty positives and repelling the feature representations from faulty negatives. We next propose solutions to these two problems.

4.1. Weighted xID: Tackling Faulty Positives

To reduce the impact of faulty positives, we propose to optimize a weighted loss. Let $w_i \in [0, 1]$ be a set of sample weights that identify faulty positives. Robustness is achieved by re-weighting the xID loss of Equation 1

$$\mathcal{L}_{\text{RxID}} = \frac{\sum_i w_i \mathcal{L}_{\text{xID}}(\mathbf{v}_i, \mathbf{a}_i)}{\sum_i w_i}. \tag{5}$$

To estimate sample weights w_i, we leverage observations from Figure 3. Since low similarities $\bar{\mathbf{v}}_i^T\bar{\mathbf{a}}_i$ are indicative of faulty positives, we define the weights w_i to be proportional to the cumulative distribution of these scores. We assume the scores to be normally distributed and define w_i as

$$w_i = t_{w_{\min}}\left(C_{\mathcal{N}}\left(\bar{\mathbf{a}}_i^T\bar{\mathbf{v}}_i; \mu + \delta\sigma, \kappa\sigma^2\right)\right), \tag{6}$$

where μ and σ^2 are the sample mean and variance of the scores, $C_{\mathcal{N}}$ is the cumulative distribution of a transformed normal distribution $\mathcal{N}(\mu+\delta\sigma, \kappa\sigma^2)$, and $t_{w_{\min}}(x) = x\cdot(1-w_{\min}) + w_{\min}$ is a soft truncation function used to assign a non-zero weight $w_{\min}$ to low score instances. δ, κ and $w_{\min}$ are shape hyper-parameters that provide flexibility to the weight function, adjusting the location and rate of decay of the weights. Figure 5 shows how the weighting function varies with the shape hyper-parameters δ, κ and $w_{\min}$.

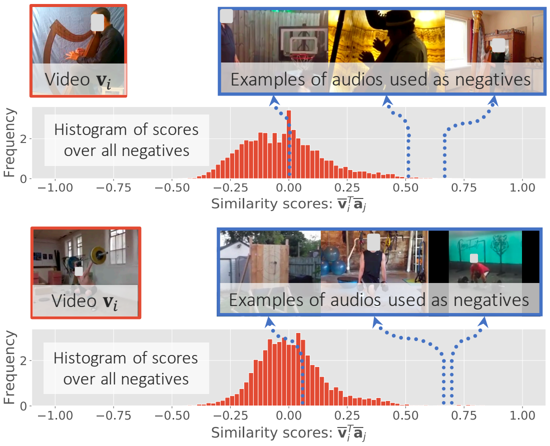

Figure 4: Faulty negatives in a pretrained cross-modal model. Two instances $\mathbf{v}_i$ and the corresponding negatives used by a xID model sorted by their similarity scores. The actual videos are provided in supplementary material. xID often uses faulty negatives for contrastive learning.

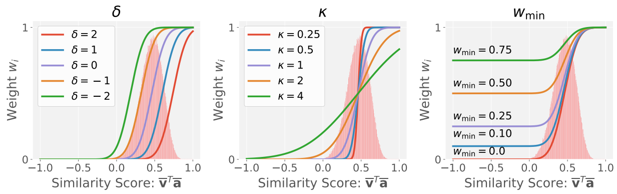

Figure 5: Weights as function of similarity scores $\bar{\mathbf{v}}_i^T\bar{\mathbf{a}}_i$ for different values of shape parameters δ, κ and $w_{\min}$. Parameters μ, σ are automatically determined from the histogram of similarity scores $\bar{\mathbf{v}}_i^T\bar{\mathbf{a}}_i$ (shown in red).

4.2. Soft Targets: Tackling Faulty Negatives

As observed in §3, faulty negatives are overemphasized during training. The underlying reason is that the xID loss of Equation 1 has too strict a definition of negatives: every negative instance $j \neq i$ is considered 'equally negative.' To limit the impact of faulty negatives, we introduce a 'softer' definition by introducing soft targets $T(j|i)$, based on the similarity between instance i and negative j. We then minimize a soft-xID loss

$$\begin{aligned}\mathcal{L}_{\text{Soft-xID}}(\mathbf{v}_i, \mathbf{a}_i) = &-\textstyle\sum_j T_v(j|i)\log P(\bar{\mathbf{a}}_j|\mathbf{v}_i;\tau)\\ &-\textstyle\sum_j T_a(j|i)\log P(\bar{\mathbf{v}}_j|\mathbf{a}_i;\tau) \quad (7)\\ T_v(j|i) = &\ (1-\lambda)\mathbf{1}_{i=j} + \lambda S_v(j|i) \quad (8)\\ T_a(j|i) = &\ (1-\lambda)\mathbf{1}_{i=j} + \lambda S_a(j|i) \quad (9)\end{aligned}$$

where $\mathbf{1}_{i=j}$ is the one-hot targets of vanilla xID, S_v and $S_a \in [0, 1]$ are softening scores (described next) used to adjust the one-hot targets, and $\lambda \in [0, 1]$ is a mixing coefficient that weighs the two terms. Equations 1 and 7 are identical when $\lambda = 0$. Since $T(j|i)$ is no longer strictly zero for similar instances, minimizing Equation 7 reduces the force to repel faulty negatives and thus their impact.

Estimating softening scores S. Since our approach focuses on self-supervised learning, we must estimate the softening scores S automatically, *i.e.*, without class labels. We describe multiple strategies for estimating these values and illustrate them in Figure 6.

- **Bootstrapping** [66] is a well established procedure to create soft targets. It uses the model's own predictions (posteriors) as the softening scores, *i.e.*,

$$S_v(j|i) = P(\bar{\mathbf{a}}_j|\bar{\mathbf{v}}_i;\tau_s) \text{ and } S_a(j|i) = P(\bar{\mathbf{v}}_j|\bar{\mathbf{a}}_i;\tau_s), \quad (10)$$

where τ_s controls the peakiness of the distribution. However, bootstrapping computes the target distribution without aggregating information from any other source other than each model's own posterior.

- **Swapped prediction** improves upon bootstrapping by using the posteriors of the opposite modality, *i.e.*, the softening scores S_v for the video modality are computed using the posterior of the audio encoder and vice-versa,

$$S_v(j|i) = P(\bar{\mathbf{v}}_j|\bar{\mathbf{a}}_i;\tau_s) \text{ and } S_a(j|i) = P(\bar{\mathbf{a}}_j|\bar{\mathbf{v}}_i;\tau_s). \quad (11)$$

As a result, in addition to the instance itself, the model is asked to predict which other instances are deemed similar in the opposite modality.

- **Neighbor prediction** relies on within-modal relationships to estimate the similarity between instances, thus avoiding potential mismatched audio and visual modalities when computing the soft targets. Specifically, we define

$$S_v(j|i) = \rho(\bar{\mathbf{v}}_i^T\bar{\mathbf{v}}_j/\tau_s) \text{ and } S_a(j|i) = \rho(\bar{\mathbf{a}}_i^T\bar{\mathbf{a}}_j/\tau_s), \quad (12)$$

where ρ is the softmax operator.

- **Cycle consistent prediction** improves upon 'swapped prediction' by focusing on negatives that are good correspondences themselves, *i.e.*, negatives with high similarity scores $\bar{\mathbf{v}}_j^T\bar{\mathbf{a}}_j$. In this case, we define

$$S_v(j|i) = \rho(\bar{\mathbf{v}}_i^T\bar{\mathbf{a}}_i/\tau_t + \bar{\mathbf{a}}_i^T\bar{\mathbf{v}}_j/\tau_s + \bar{\mathbf{v}}_j^T\bar{\mathbf{a}}_j/\tau_t) \quad (13)$$

$$S_a(j|i) = \rho(\bar{\mathbf{a}}_i^T\bar{\mathbf{v}}_i/\tau_t + \bar{\mathbf{v}}_i^T\bar{\mathbf{a}}_j/\tau_s + \bar{\mathbf{a}}_j^T\bar{\mathbf{v}}_j/\tau_t) \quad (14)$$

where τ_s and τ_t control the relative importance of swapped prediction target and avoiding negatives with weak correspondences. As shown in Figure 6, the terms $\bar{\mathbf{v}}_i^T\bar{\mathbf{a}}_i$ and $\bar{\mathbf{v}}_j^T\bar{\mathbf{a}}_j$ complete a cycle over instances i and j.

How do soft targets mitigate faulty negatives? The soft xID loss of Equation 7 prevents overemphasizing faulty negatives by relying on soft targets $T(j|i)$ that encode similarities between instances. To better understand the mechanism, we examine the partial derivatives of the soft-xID loss:

$$-\frac{\partial L_{\text{Soft-xID}}}{\partial \mathbf{v}_i} = \sum_j \frac{\bar{\mathbf{a}}_j}{\tau}(T_v(j|i) - P(\bar{\mathbf{a}}_i|\mathbf{v}_i)) \quad (15)$$

$$-\frac{\partial L_{\text{Soft-xID}}}{\partial \mathbf{a}_i} = \sum_j \frac{\bar{\mathbf{v}}_j}{\tau}(T_a(j|i) - P(\bar{\mathbf{v}}_i|\mathbf{a}_i)). \quad (16)$$

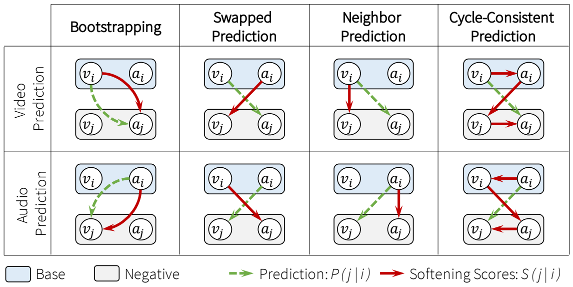

Figure 6: Strategies to estimate softening scores $S(i|j)$.

Since faulty negatives j tend to be similar to the base instance i, the soft targets $T(j|i)$ are higher. Thus, the target representations $\bar{\mathbf{v}}_j$ and $\bar{\mathbf{a}}_j$ of faulty negatives act as weaker negatives, or even as positives when $T(j|i)$ is larger than the model posteriors.

4.3. Training

We introduced two procedures to deal with noisy training signals inherent to cross-modal instance discrimination. §4.1 presents a weighting mechanism that limits the effect of faulty positives, while §4.2 proposes a soft instance discrimination loss that predicts relations between instances, thus preventing the training algorithm from overemphasizing faulty negatives. Since both procedures rely on the alignment between audio and visual target representations to find weak correspondences, we start by training the model for cross-modal instance discrimination alone (Equation 1). After the initial warmup stage, the two procedures can be combined by minimizing

$$\mathcal{L} = \frac{1}{\sum_k w_k} \sum_i w_i \mathcal{L}_{\text{Soft-xID}}(\mathbf{v}_i, \mathbf{a}_i) \quad (17)$$

where w_i are the sample weights of Equation 6 and $\mathcal{L}_{\text{Soft-xID}}$ is the xID loss with soft targets of Equation 7.

5. Experiments

We perform experiments to better understand cross-modal learning and validate the proposed improvements. We pretrain models on a subset of the Kinetics-400 [76] dataset containing 50K videos and evaluate the pretrained models by transfer learning.

5.1. Experimental Setup

Video and audio preprocessing. During training, we extract video clips of length $T = 8$ frames and resolution 80×80 at 16 fps. Video clips are augmented using temporal jittering, multi-scale cropping, horizontal flipping, color jittering, gray-scaling, and Gaussian blur [14]. All data augmentations are applied consistently over all frames. For the audio, we extract mono clips of length 2s at a sample rate of 11025Hz, and compute log spectrograms on 50ms windows with a hop size of 25ms. The spectrogram is then converted to a mel scale with 80 bands, yielding an audio input of size 80×80. Audio data is augmented by randomly changing the volume by at most 20%.

Video and audio models. The video encoder is a 9-layer version of the R(2+1)D model of [75]. Following [5, 56], we replaced global average pooling with max pooling. The audio encoder is a 9-layer 2D ConvNet with batch normalization and global max pooling. Both encoders yield 512-dimensional features, which are mapped into a 128-dimensional sphere using a non-linear projection head (as in [14]) followed by L2 normalization.

Pretraining. In the warm-up stage, the video and audio models are trained to optimize the loss of Equation 1 using the Adam optimizer [40] with default hyper-parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) for 400 epochs with a learning rate of $1e-4$ and a batch size of 224 split over 2 12Gb GPUs. In order to reduce the memory footprint of our models, we employ mixed-precision training [50] using PyTorch AMP [59]. Following [56, 81], the audio and video target representations, $\bar{\mathbf{a}}$ and $\bar{\mathbf{v}}$, are generated using memory banks updated by exponential moving average with an update constant of 0.5. The contrastive loss of Equation 1 is defined by opposing the target representation of the opposite modality to 1024 negatives randomly drawn from the memory bank. The temperature hyper-parameter is set to $\tau = 0.07$.

After the initial warm-up stage, models are trained for an additional 200 epochs to optimize the loss of Equation 17 using the Adam optimizer and a cosine learning rate schedule starting at $1e-4$ and ending at $1e-5$. The hyper-parameters for the weighting function (Equation 6) and the soft xID loss (Equation 7) are discussed below. To provide a fair comparison to the AVID baseline [56], we control for the number of epochs by training the baseline model for an additional 200 epochs as well.

Downstream tasks. We evaluate audio and video features using transfer learning. Video features are evaluated on the UCF [72] and HMDB [42] datasets. Models are fine-tuned using 8-frame clips for 200 epochs using the Adam optimizer with a batch size of 192 on a single GPU and a cosine learning rate schedule starting at $1e-4$ and ending at $1e-5$. To prevent overfitting, we use dropout after the global max-pooling layer, weight decay of $1e-3$, and reduced the learning rate for backbone weights by a factor of 10. At test time, top-1 accuracy is measured on video level predictions computed by averaging the predictions of 10 clips uniformly sampled over the entire video.

Following [8, 83], we also evaluate the quality of video representations by conducting retrieval experiments without fine-tuning. Feature maps of size $4{\times}4{\times}512$ are extracted from 10 clips per video and averaged. We then use videos in the test set to query the training set. As in [8, 83], a correct retrieval occurs when the class of one of the top-k retrieved videos matches the query, and performance is measured by the average top-k retrieval performance ($R@K$).

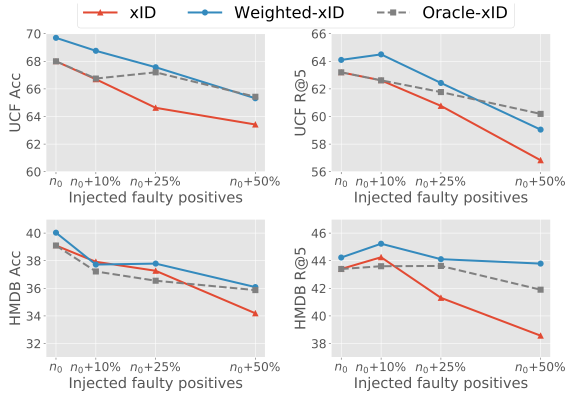

Figure 7: Faulty positives *vs*. transfer performance We inject faulty positives in the pre-training data (Kinetics) by randomly replacing the audio corresponding to a video. We evaluate the pretrained models on action recognition and see that increasing the fraction of faulty positives hurts transfer performance. Our weighted loss (Weighted-xID) is less sensitive to faulty positives and even outperforms an oracle version that information about altered samples. This is because the unaltered pretraining dataset itself has an unknown, but significant number (n_0) of faulty positives.

5.2. Weighted cross-modal learning

We analyze the impact of faulty positives on the representations learned by cross-modal instance discrimination.

Faulty positives are detrimental to representation learning. We artificially control the number of faulty positives to assess their impact on representation learning. The pretraining dataset *already contains* an unknown (but significant) number of faulty positives. We increase this number by injecting more faulty positives. A faulty positive is injected by replacing the audio of an instance with a randomly selected audio that is not part of the training set. After pretraining, the learned visual representation is evaluated on the UCF and HMDB datasets using both classification and retrieval protocols. Figure 7 shows that as the fraction of faulty positives increases, the transfer performance of cross-modal instance discrimination (xID) decreases significantly.

Weighted xID reduces the impact of faulty positives. We evaluate the effectiveness of the weighted xID loss (Equation 5) as a function of the number of faulty positives. We compare the representations learned by Weighted-xID to its unweighted counterpart (xID), as well as an oracle weight function (Oracle-xID) which assigns $w_i = 0$ to artificially altered instances and $w_i = 1$ otherwise. The weight function of Equation 5 is defined with $\kappa = 0.5$ and $w_{\text{min}} = 0.25$. For simplicity, we assume that the noise level is known and set δ in Weighted-xID so that the midpoint of the weighting function coincides with the known fraction of altered samples. In practice, the noise level would need to be estimated either by cross-validation or by manual inspection. Weighted-xID is not very sensitive to these parameters (see appendix).

Figure 7 shows the performance of the three approaches. Oracle-xID consistently outperforms xID when the fraction

Table 1: Different strategies for computing soft targets in the pretraining loss of Equation 7. Models are pretrained on Kinetics and evaluated on UCF and HMDB datasets using fine-tuning and retrieval protocols. Best method is bolded. Second best is underlined.

Target Distribution	UCF		HMDB	
	Acc	R@5	Acc	R@5
Oracle*	73.6	76.0	45.4	53.6
xID [56]	68.0	63.2	39.0	43.4
Bootstrapping	69.2	64.4	40.5	44.7
Neighbor Pred.	**70.5**	65.4	41.2	45.0
Swapped Pred.	70.0	64.9	41.3	45.4
CCP	70.3	**65.9**	**41.5**	**45.5**

**Uses class labels to generate target distribution.*

of injected faulty positives is high. This shows that the detrimental impact of noisy correspondences can be mitigated with a weighting strategy. Weighted-xID also outperforms the unweighted version (xID) in nearly all cases, with larger margins for larger fractions of noisy correspondences. In fact, Weighted-xID even outperforms the oracle weight function, especially at lower noise levels. This is because the original Kinetics dataset *already contains* a significant amount of weak correspondences, which the oracle weight function treats as clean $w_i = 1$, while the weighting function of Equation 6 can suppress them.

5.3. Instance discrimination with soft targets

To limit the impact of faulty negatives, we proposed to match a soft target distribution that encodes instance similarity. We analyze different design decisions for creating the soft targets and their effect on transfer performance.

Comparison of strategies for computing targets As summarized in Figure 6, the soft target distributions can be computed by aggregating evidence from all modalities. Four different strategies were proposed, bootstrapping, swapped or cycle consistent assignments. Models were trained to minimize the loss of Equation 7 with $\lambda = 0.5$. We empirically found that peakier target distributions work better, and set the temperature parameter τ_s to 0.02. For cycle consistent assignments, the terms $\bar{\mathbf{v}}_j^T \bar{\mathbf{a}}_j$ are used so as to focus on negatives that are good correspondences themselves. A temperature hyper-parameter of $\tau_t = 0.07$ was sufficient to impose such constraint. Beyond the baseline xID, we also compare to an *oracle* target distribution that has access to class labels to determine the similarity between instances. Specifically, the oracle considers two instances i and j to be similar if they share the same class label, and computes $T_v(j|i)$ and $T_a(j|i)$ by assigning a uniform distribution over similar instances, and 0 to non-similar ones.

Table 1 shows the performance of different target distributions. We observe a large gap between vanilla xID and xID with an oracle soft target, which demonstrates the detrimental effect of faulty negatives. In the self-supervised case, however, labels are not available for determining the target distribution. Nevertheless, the estimated target distributions (bottom four rows) still significantly improve over the xID loss. Regarding the various types of target distributions, bootstrapping is the least effective. This is expected since, in this case, the target distribution is a peakier version of the model posterior, *i.e.* it is obtained without aggregating information from any other sources. Cycle consistent prediction is the most effective most often. This is because cycle consistent prediction not only leverages the opposite modality to create the target distribution, but it also avoids targets that are not good correspondences themselves, *i.e.*, avoids samples with low cross-modal similarities.

Table 2: Combining weighted xID loss with soft targets. Models are pretrained on Kinetics with the loss of Equation 17 and evaluated on UCF and HMDB datasets using fine-tuning and retrieval protocols.

Method	Robust Weighting	CCP Soft Targets	UCF		HMDB	
			Acc	R@5	Acc	R@5
xID [56]	✗	✗	68.0	63.2	39.0	43.4
Weighted-xID	✓	✗	69.7	64.1	40.1	44.3
Soft-xID	✗	✓	70.3	65.9	41.5	45.5
Robust-xID	✓	✓	**71.6**	**67.4**	**41.9**	**46.2**

5.4. Robust instance discrimination with soft targets

Sample weighting and soft targets are designed to address two different sources of noisy training signals inherent to cross-modal contrastive learning: faulty positives and faulty negatives. Table 2 shows that the two proposed improvements (Weighted-xID and Soft-xID) not only improve upon the representations of vanilla xID, they are also complementary to each other. By combining the two approaches using the loss of Equation 17, Robust-xID improved upon Weighted and Soft-xID.

6. Comparison to prior work

We compare Robust-xID to prior work in self-supervised learning. We train our models on the Kinetics dataset, using an 18-layer R(2+1)D model [75] for the video, and a 9-layer 2D ConvNet with batch normalization for the audio. Video clips of length 8-frames and 112×112 resolution are extracted at 16fps, and the same data augmentations from §5 are used. We extract audio clips of length 2s at 24KHz and compute log mel spectrograms with 128 time steps and 128 frequency bands. All models are trained with the Adam optimizer with a batch size of 512 distributed across 8 12Gb GPUs. We warm-up the models for 200 epochs by training on the xID loss alone with a learning rate of $5e-4$. The models are then trained with sample weights and cycle consistent soft targets for an additional 200 epochs using a cosine learning rate schedule from $5e-4$ to $5e-5$.

After pre-training, models are evaluated on UCF and HMDB. We fine-tune the models using either 8 or 32 frame clips for action recognition and report the top-1 accuracy of video level predictions (with 10 clips per video) in Ta-

Table 3: Comparison to prior work (finetuning). Performance on the downstream UCF and HMDB datasets by full network fine-tuning after pre-training on Kinetics. We report top-1 accuracy of video level predictions (10 clips per video). We also list the video encoder, amount of compute used for pre-training and the fine-tuning resolution.

Method	Model	Compute # GPUs	Finetuning Resolution	UCF	HMDB
DPC [33]	S3D	4	25×128^2	75.7	35.7
CBT [73]	S3D	8	16×112^2	79.5	44.6
Multisensory [58]	3D-ResNet18	3	32×224^2	82.1	–
AVTS [41]	MC3-18	4	25×224^2	84.1	52.5
SeLaVi [8]	R(2+1)D-18	64	32×112^2	83.1*	47.1*
XDC [3]	R(2+1)D-18	64	8×224^2	74.2*	39.0*
	R(2+1)D-18	64	32×224^2	86.8*	52.6*
AVID-CMA [56]	R(2+1)D-18	64	8×224^2	83.7*	49.5*
	R(2+1)D-18	64	32×224^2	87.5*	60.8*
GDT [61]	R(2+1)D-18	64	32×224^2	89.3*	60.0*
xID+CMA [56]	R(2+1)D-18	8	8×112^2	80.6	48.6
	R(2+1)D-18	8	32×112^2	84.9	54.7
Robust-xID	R(2+1)D-18	8	8×112^2	81.9	49.5
	R(2+1)D-18	8	32×112^2	**85.6**	**55.0**

* *Models pre-trained with more than one compute node (8 GPUs).*

Table 4: Retrieval performance on UCF and HMDB datasets after pre-training on Kinetics for different numbers of retried neighbors.

Method	UCF			HMDB		
	R@1	R@5	R@20	R@1	R@5	R@20
SpeedNet [9]	13.0	28.1	49.5	-	-	-
VCP [48]	18.6	33.6	53.5	7.6	24.4	53.6
VSP [16]	24.6	41.9	76.9	10.3	26.6	54.6
CoCLR [35]	55.9	70.8	82.5	26.1	45.8	69.7
SeLaVi [8]	52.0	68.6	84.5	24.8	47.6	75.5
GDT [61]	57.4	73.4	88.1	25.4	51.4	75.0
xID+CMA [56]	60.1	76.6	90.1	29.7	54.4	77.1
Robust-xID	**60.9**	**79.4**	**90.8**	**30.8**	**55.8**	**79.7**

ble 3. The proposed procedure outperformed all prior work where pretraining is limited to a single node (8 GPUs), and even outperformed methods like SeLaVi, which require 8× more compute for training. We also conducted a close comparison to the CMA procedure of [56] (xID+CMA). While CMA can also partially address the problem of faulty negatives, Robust-xID showed better performance. Robust-xID is also easier to implement as it identifies both faulty positives and negatives in a simpler online fashion. We note that xID+CMA is a faithful implementation of AVID+CMA [56], as it follows the original code with improved data augmentations. However, the results reported for xID+CMA are lower than those originally reported in [56] because 1) distributed training was conducted on 8 GPUs instead of 64 (large batch sizes are known to have a substantial impact on contrastive learning performance [14, 15, 11]), and 2) [56] is trained and evaluated with videos of higher resolution (224 instead of 112). By training the proposed model with a larger batch size, we expect the performance to improve further.

We also compare the learned representations to prior work without fine-tuning. Following [8, 61], we conducted retrieval experiments, and report the retrieval performance $R@K$ for $K = 1$, $K = 5$ and $K = 20$ neighbors in Table 4. The retrieval protocol was described in §5. Following [38, 61], we also assessed the few-shot learning performance of Robust-xID models on UCF and HMDB. For the few-shot evaluation, we average the pretrained max-pooling features of 10 clips per video. The features from n videos per class are then used to learn a one-vs-all linear SVM classifier with $C = 1$. We report the top-1 accuracy averaged over 50 trials in Table 5. On both the retrieval and few-shot learning tasks, Robust-xID improves significantly over all prior work, reaffirming the importance of mitigating the training noise introduced by faulty positives and faulty negatives.

Table 5: Few-shot learning on UCF and HMDB after pre-training on Kinetics. Classification is conducted using a one-vs-all SVM trained on the pretrained features of n images per class. Top-1 accuracy is reported for $n \in \{1, 5, 20\}$.

Method	UCF			HMDB		
	1-shot	5-shot	20-shot	1-shot	5-shot	20-shot
3D-RotNet [38]	15.0	31.5	47.1	-	-	-
GDT [61]	26.3	42.4	49.4	13.4	15.6	20.8
xID+CMA [56]	30.8	53.1	66.9	13.5	25.0	33.6
Robust-xID	**32.8**	**54.6**	**67.8**	**14.1**	**25.9**	**34.9**

7. Discussion and future work

We identified and tackled two significant sources of noisy training signals in audio-visual instance discrimination, namely instances with weak audio-visual correspondence (or faulty positives) and semantically similar negatives (or faulty negatives). We showed the impact of faulty correspondences on representation learning by removing them using an oracle with access to ground-truth annotations. We then proposed a method that mitigates the impact of faulty correspondences without relying on ground-truth annotations. Extensive analysis and experimental evaluations show that the proposed procedure enhances representation learning and improves transfer performance significantly.

Our findings show that cross-modal learning should be seen as a problem of learning with noisy targets. While we propose two specific methods to address faulty positives and faulty negatives (*i.e.* weighting and soft targets), there is a rich literature regarding supervised learning with noisy labels. Developing methods that tackle noisy correspondences are a promising avenue for future research. Furthermore, we focused on audio-visual learning, but other pairs of modalities such as RGB and flow or text from instructional videos also present similar problems. We believe that our method will also benefit cross-modal learning from other modalities.

Acknowledgements This work was partially funded by NSF awards IIS-1924937, IIS-2041009, and gifts from Amazon and Qualcomm. We also acknowledge and thank the use of the Nautilus platform for some of the experiments above.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Deep lip reading: A comparison of models and an online application. *Proc. Interspeech*, 2018. 2

[2] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-Supervised MultiModal Versatile Networks. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2020. 1

[3] Humam Alwassel, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Adv. Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 8

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Eur. Conf. Computer Vision (ECCV)*, 2018. 2

[5] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. *Int. Conf. Computer Vision (ICCV)*, 2017. 1, 2, 6

[6] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin Mcguinness. Unsupervised label noise modeling and loss correction. In *Int. Conf. on Machine Learning (ICML)*, 2019. 3

[7] YM Asano, C Rupprecht, and A Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *Int. Conf. on Machine Learning (ICML)*, 2019. 2

[8] Yuki M. Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2020. 2, 6, 8

[9] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 8

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Eur. Conf. Computer Vision (ECCV)*, 2018. 2

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 3, 8

[12] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2017. 3

[13] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Eur. Conf. Computer Vision (ECCV)*, 2020. 2

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Int. Conf. on Machine Learning (ICML)*, 2020. 1, 2, 3, 5, 6, 8

[15] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 8

[16] Hyeon Cho, Taehoon Kim, Hyung Jin Chang, and Wonjun Hwang. Self-supervised spatio-temporal representation learning using variable playback speed prediction. *arXiv preprint arXiv:2003.02692*, 2020. 8

[17] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

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

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

[20] Aditya Deshpande, Jason Rock, and David Forsyth. Learning large-scale automatic image colorization. In *Int. Conf. Computer Vision (ICCV)*, 2015. 2

[21] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Int. Conf. Computer Vision (ICCV)*, 2015. 2

[22] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2014. 1, 2

[23] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[24] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2013. 3

[25] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[26] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Eur. Conf. Computer Vision (ECCV)*, 2018. 2

[27] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[28] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[29] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI Conference on Artificial Intelligence*, 2017. 3

[30] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Int. Conf. Learning Representations (ICLR)*, 2018. 2

[31] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2006. 1, 3

[32] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2018. 3

[33] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Eur. Conf. Computer Vision Workshops (ECCV-W)*, 2019. 2, 8

[34] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *Eur. Conf. Computer Vision (ECCV)*, 2020. 2

[35] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2020. 2, 8

[36] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3

[37] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2018. 3

[38] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018. 8

[39] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI Conference on Artificial Intelligence*, 2019. 2

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

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

[42] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *Int. Conf. Computer Vision (ICCV)*. IEEE, 2011. 6

[43] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *Eur. Conf. Computer Vision (ECCV)*, 2016. 2

[44] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[45] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[46] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[47] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Int. Conf. Computer Vision (ICCV)*, 2017. 3

[48] Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. Video cloze procedure for self-supervised spatio-temporal learning. In *AAAI Conference on Artificial Intelligence*, 2020. 8

[49] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *Int. Conf. on Machine Learning (ICML)*, 2018. 3

[50] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017. 6

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

[52] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Eur. Conf. Computer Vision (ECCV)*, 2016. 2

[53] Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In *Int. Conf. on Machine Learning (ICML)*, 2009. 2

[54] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2020. 2

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

[56] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 6, 7, 8

[57] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Eur. Conf. Computer Vision (ECCV)*, 2016. 2

[58] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Eur. Conf. Computer Vision (ECCV)*, 2018. 1, 2, 8

[59] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Information Processing Systems (NeurIPS)*. 2019. 6

[60] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[61] Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal

self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020. 1, 2, 3, 8

[62] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[63] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 3

[64] AJ Piergiovanni, Anelia Angelova, and Michael S. Ryoo. Evolving losses for unsupervised video representation learning. *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[65] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[66] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014. 3, 5

[67] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Int. Conf. on Machine Learning (ICML)*, 2018. 3

[68] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[69] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2019. 3

[70] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *Int. Conf. on Machine Learning (ICML)*, 2019. 3

[71] Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*, 2020. 3

[72] 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. 6

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

[74] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Eur. Conf. Computer Vision (ECCV)*, 2020. 1, 2, 3

[75] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018. 6, 7

[76] 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. 5

[77] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *Eur. Conf. Computer Vision (ECCV)*, 2020. 2

[78] Xudong Wang, Ziwei Liu, and Stella X Yu. Unsupervised feature learning by cross-level discrimination between instances and groups. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 3

[79] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Int. Conf. Computer Vision (ICCV)*, 2019. 3

[80] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[81] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3, 6

[82] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Delving into inter-image invariance for unsupervised visual representations. *arXiv preprint arXiv:2008.11702*, 2020. 2

[83] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[84] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *Int. Conf. on Machine Learning (ICML)*, 2017. 3

[85] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *Int. Conf. Learning Representations (ICLR)*, 2018. 3

[86] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Eur. Conf. Computer Vision (ECCV)*, 2016. 2

[87] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[88] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Adv. Neural Information Processing Systems (NeurIPS)*, 2018. 3

[89] Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[90] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Int. Conf. Computer Vision (ICCV)*, 2019. 2

[91] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Eur. Conf. Computer Vision (ECCV)*, 2018. 2

A. Parametric studies

We provide a parametric study of key Robust-xID hyper-parameters.

Weight function shape parameter δ One critical parameter of Weighted-xID is the shape parameter δ, which specifies the mid-point location of the weight function. For example, when $\delta = -2$, the midpoint is located at $\mu - 2\sigma$ where μ and σ are the sample mean and standard deviation of the scores $\bar{\mathbf{v}}_i^T \bar{\mathbf{a}}_i$. This means that for $\delta = -2$, the majority of samples will have a weight of 1, and only a small fraction will have a weight close to $w_{\min}$. As δ increases, the proportion of samples that are down-weighted also increases. To study the impact of δ, we trained several models using Weighted-xID with different values of δ and for different amounts of injected faulty positives n_0. Other hyper-parameters were kept at their default values $w_{\min} = 0.25$ and $\kappa = 0.5$. The transfer performance is shown in Figure 8. As can be seen, the proposed robust xID procedure is not very sensitive to this hyper-parameter. This suggests that Robust-xID can help representation learning as long as clear faulty positives are suppressed.

Soft-xID: Mixing coefficient The mixing coefficient λ specifies the degree to which the one-hot targets of instance discrimination are softened in Soft-xID. The one-hot instance discrimination targets are used when $\lambda = 0$. As λ increases, the softening scores $S(j|i)$ are increasingly used to adjust the one-hot targets. To study the impact of the mixing coefficient λ, we trained several models using Soft-xID with various values of λ. Cycle consistent targets were used as the softening strategy. Figure 9 shows the transfer performance of the learned models on UCF and HMDB under the fine-tuning and retrieval protocols. The trend is consistent across the two datasets and two evaluation protocols. Softening the instance discrimination targets enhances representation learning, with the optimal performance achieved with a mixing coefficient between 0.25 and 0.5. However, as the mixing coefficient increases substantially $\lambda > 0.65$, the targets are derived from the model prediction alone and disregard instance labels. In this case of large λ, the pre-training fails completely, *i.e.*, the learned representations have very low transfer performance.

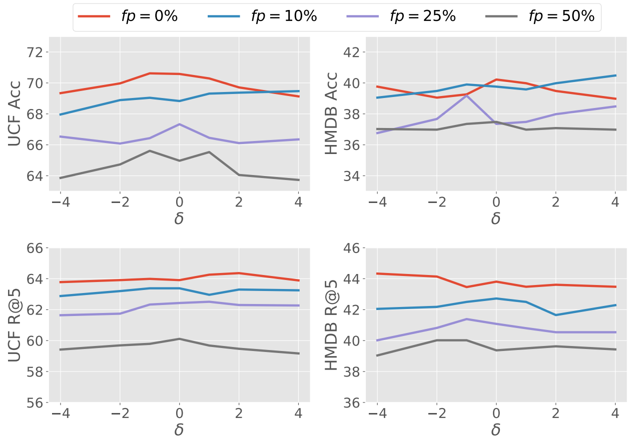

Figure 8: Effect of shape parameter δ in Weighted-xID. Transfer learning performance is evaluated on two datasets (UCF and HMDB) under two protocols (full finetuning and retrieval). For the fine-tuning protocol, we report final accuracy of video level predictions. For the retrieval protocol, we report $R@5$.

B. Additional analysis

The proposed approach learns high-quality feature representations that can be used to discriminate several action classes. This was shown in the main paper by reporting transfer learning results. We now provide additional qualitative evidence and analysis.

Retrieval For each video, we extracted $4 \times 4 \times 512$ feature maps from the video encoder learned using Robust-xID on the full Kinetics dataset. Figure 11 depicts the top 4 closest videos for several query samples. As can be seen, Robust-xID produces highly semantic features, enabling correct retrievals for a large number of videos spanning a large number of classes. Furthermore, even when a video of a different class is retrieved, the errors are intuitive (for example, the confusion between 'American football' and 'Hurling' in the third row). Failure cases also seem to be correlated with classes that are hard to distinguish from the audio alone (eg, different types of kicking sports or swimming strokes).

Class-based analysis To better understand which classes are better modeled by the Robust-xID framework, we measured the top-1 retrieval performance ($R@1$) averaged across all images of each class. Similar to the analysis above, each video is represented by a $4 \times 4 \times 512$ feature map extracted from a video encoder learned using Robust-xID on the full Kinetics dataset. Figure 10 depicts a list of Kinetics classes sorted by their average $R@1$ score. As can be seen, action classes which are often accompanied by long and distinctive sounds (*e.g.*, squash, harp, drums, accordion, or scuba

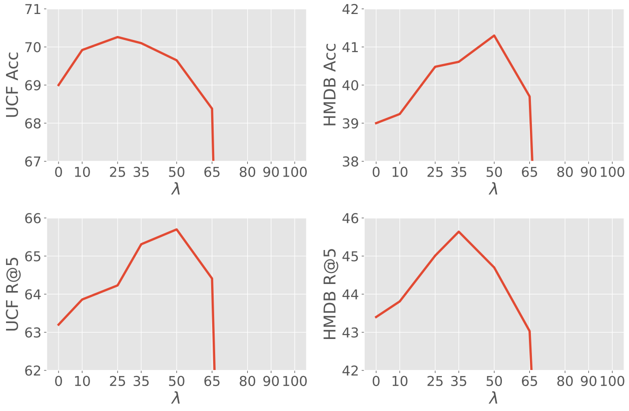

Figure 9: Effect of mixing coefficient λ in Soft-xID. Transfer learning performance is evaluated on two datasets (UCF and HMDB) under two protocols (full finetuning and retrieval). For the fine-tuning protocol, we report final accuracy of video level predictions. For the retrieval protocol, we report $R@5$.

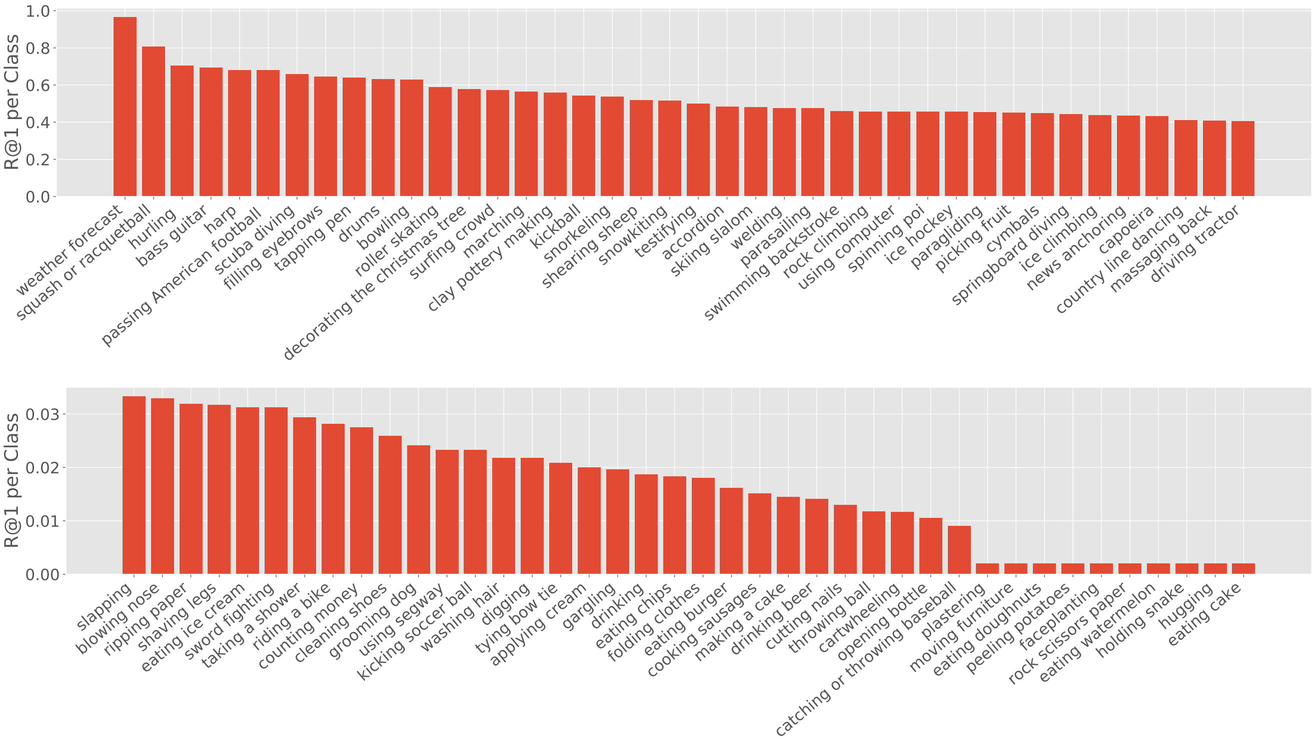

Figure 10: Best and worse Kinetics classes. For each class, we depict the top-1 retrieval performance ($R@1$) averaged across all images of each class. The plot above shows the top 40 classes and the plot below the bottom 40 classes.

diving) tend to be more easily distinguished from others. In contrast, classes with less distinctive audio (*e.g.*, making a cake, eating cake, or hugging) or classes where distinctive sounds are short-lived (*e.g.*, blowing nose, gargling or kicking ball) are harder to model using a cross-modal audio-visual framework. As a result, the features learned for such classes are less discriminative.

Faulty positive detection performance To obtain a rough estimate of performance of the faulty positive detection procedure, we randomly sampled 100 videos from the 10000 most likely faulty positives, as identified by Robust-xID trained on the full Kinetics dataset. We then manually labeled them according to how related their audio and visual signals are. From those, 67 were clear faulty pairs; 24 contained narrative voice-overs (*i.e.*, required natural language understanding to link the two modalities); and 9 samples were clearly misidentified.

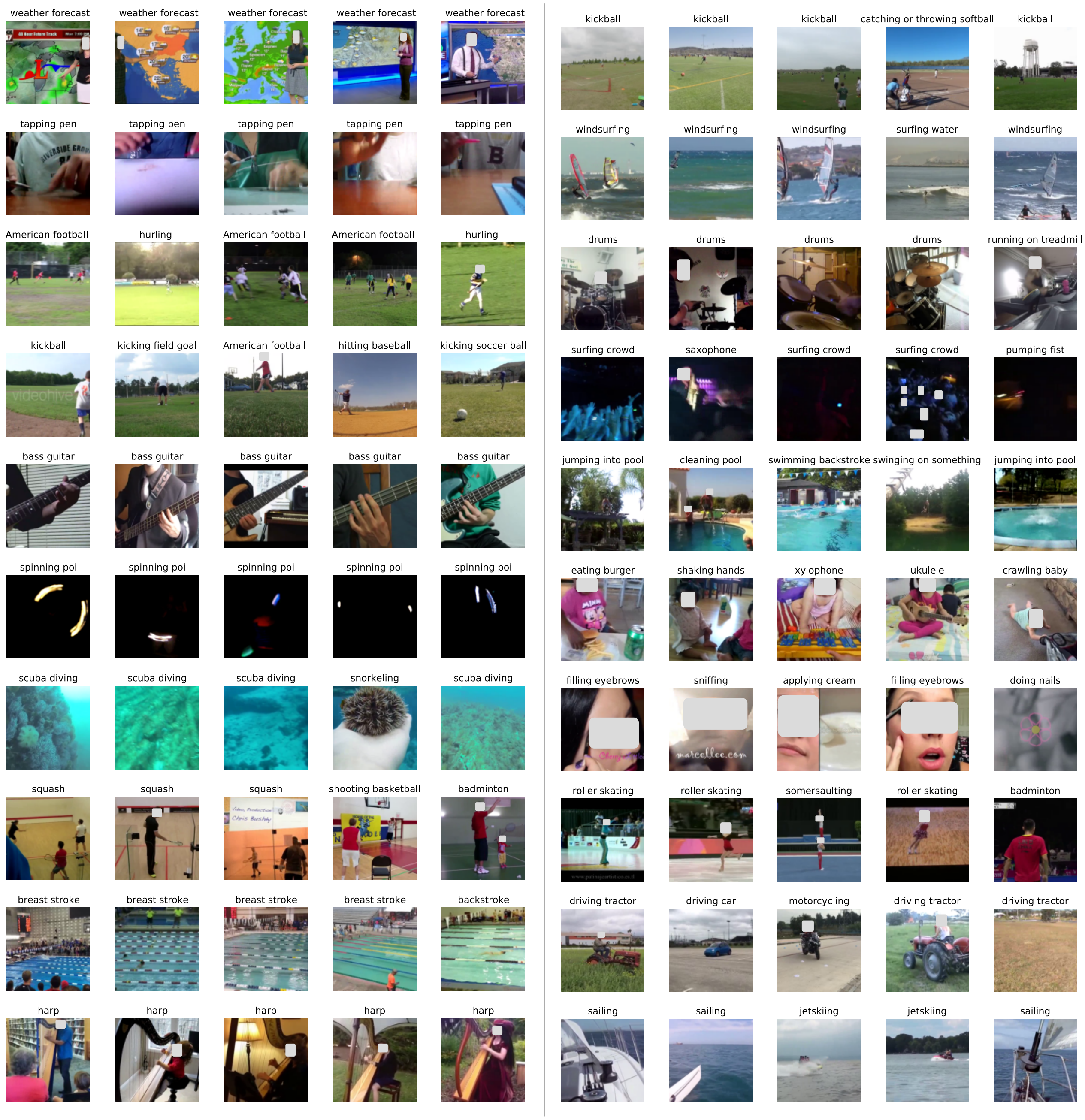

Figure 11: Retrievals. In each row, the first image depicts the query video, and the following four images depict the top 4 retrievals. The corresponding Kinetics class description is provided above each frame. Each video is represented by a $4 \times 4 \times 512$ feature map produced by the video encoder learned using Robust-xID on the full Kinetics dataset. Euclidean distance is used to determine video similarity.