Robust Audio-Visual Instance Discrimination via Active Contrastive Set Mining

Hanyu Xuan¹, Yihong Xu², Shuo Chen³, Zhiliang Wu¹, Jian Yang¹, Yan Yan⁴*, Xavier Alameda-Pineda²

¹School of Computer Science and Engineering, Nanjing University of Science and Technology, China
²Inria, University Grenoble Alpes, CNRS, Grenoble INP, LJK, Grenoble, France
³RIKEN Center for Advanced Intelligence Project, Japan
⁴Department of Computer Science, Illinois Institute of Technology, USA

{xuanhanyu, wu_zhiliang, csjyang}@njust.edu.cn, {yihong_xu, xavier.alameda-pineda}@inria.fr, shuo.chen.ya@riken.jp, yyan34@iit.edu

Abstract

The recent success of audio-visual representation learning can be largely attributed to their pervasive property of audio-visual synchronization, which can be used as self-annotated supervision. As a state-of-the-art solution, Audio-Visual Instance Discrimination (AVID) extends instance discrimination to the audio-visual realm. Existing AVID methods construct the contrastive set by random sampling based on the assumption that the audio and visual clips from all other videos are not semantically related. We argue that this assumption is rough, since the resulting contrastive sets have a large number of faulty negatives. In this paper, we overcome this limitation by proposing a novel Active Contrastive Set Mining (ACSM) that aims to mine the contrastive sets with informative and diverse negatives for robust AVID. Moreover, we also integrate a semantically-aware hard-sample mining strategy into our ACSM. The proposed ACSM is implemented into two most recent state-of-the-art AVID methods and significantly improves their performance. Extensive experiments conducted on both action and sound recognition on multiple datasets show the remarkably improved performance of our method.

1 Introduction

Visual scenes are typically synchronised with a mixture of sounds, for instance moving lips and speech or moving cars and engine noise [Xuan et al., 2020]. This pervasive audio-visual synchronization comes from the fact that sound is produced by the vibration of objects. Through such an audio-visual synchronised perception [Smith and Gasser, 2005], humans are capable of developing skills to better perceive the world. As for machine intelligence, audio-visual synchronization in videos also raises the possibility to develop similar skills by investigating audio-visual representation learning. In addition, different from expensive and subjective manual annotations, audio-visual synchronization provides free and extensive self-annotated supervisions for exploring audio-visual representation learning using numerous videos available on the Internet.

Recent studies on self-supervised audio-visual representation learning [Arandjelovic and Zisserman, 2017; Arandjelovic and Zisserman, 2018; Korbar et al., 2018; Owens and Efros, 2018] set up a binary verification problem that predicts whether an input audio-visual pair is correct, that is an “in-sync” video and audio, or incorrect, that is constructed by using “out-of-sync” audio [Korbar et al., 2018] or audio from a different video [Arandjelovic and Zisserman, 2017]. However, exploiting one single pair at a time loses the key opportunity to reason about the data distribution in large-scale.

To alleviate the above issue, some works have attempted to improve upon the binary verification problem by posing it as an Instance Discrimination (ID) [Jaiswal et al., 2021] task, called Audio-Visual Instance Discrimination (AVID). Specifically, [Morgado et al., 2021a; Morgado et al., 2021b] propose a cross-modal version of ID and [Ma et al., 2021] present a multi-modal extension of Momentum Contrast (MoCo) [He et al., 2020]. They have achieved impressive performance on AVID through constructing a contrastive set with large negatives by randomly sampling from a memory bank [Wu et al., 2018] or a queue-based dictionary [He et al., 2020]. The reasons for the success are two-fold: on one hand, the introduction of a contrastive set rather than a single pair facilitates rea-
soning about the data distribution. On the other hand, memory bank and dictionary decouple the size of the contrastive set from the mini-batch size, allowing it to be large. Theoretically, a large contrastive set can achieve a tighter lower bound on Mutual Information (MI) [McAllester and Stratos, 2020].

However, these methods rely on one key assumption when constructing the contrastive set for AVID: the audio and visual clips from all other videos are not semantically related. In practice, such an assumption does not hold for a significant amount of real-world videos. Simply increasing the size of the contrastive set beyond a threshold does not improve or even harm the performance of learned representations on downstream tasks [Saunshi et al., 2019]. We argue that such a contrastive set, constructed by randomly sampling rather than carefully designing, is much responsible for this: as shown in Fig.1, it can lead to faulty negatives (semantically related to the anchor). The contrastive set with such faulty negatives undermines the primary goal of AVID, i.e., we cannot keep the audio-visual pairs with related semantics both far from and close to each other in the representation space. Also, this issue can get aggravated as the size of the contrastive set increases.

For this purpose, we propose a method called Active Contrastive Set Mining (ACSM) that aims to mine the contrastive sets with informative and diverse negatives for robust AVID. Specifically, the semantic structures are explicitly imposed to unlabelled audio-visual clips to encourage learning a “semantics-aware” discriminative space. At this point, the contrastive set can be constructed under the guidance of progressive semantics during the training of our ACSM. Different from the existing AVID [Morgado et al., 2021a; Morgado et al., 2021b; Ma et al., 2021] which pull away each audio or visual clip from all other videos in the representation space, our ACSM takes also the diverse semantics into account, i.e., only the audio-visual pairs with diverse semantics are far away from each other. Moreover, benefiting from our semantics-aware beyond instance-specific AVID, we introduce semantic ambiguity, rather than simple time gaps [Bruno et al., 2018], to mine hard samples, which can be easily integrated into AVID. In summary, the main contributions of this work are re-emphasised as follows:

- An active contrastive set mining (ACSM) is proposed for mining the contrastive sets with informative and diverse negatives for robust AVID;
- We elegantly extend our ACSM to two state-of-the-art AVID methods, i.e., Cross-modal ID [Ma et al., 2021] and Multi-modal MoCo [Morgado et al., 2021a; Morgado et al., 2021b], which then achieves significant improvements;
- With the semantics from our ACSM, a hard sample mining strategy based on the semantic ambiguity is introduced to boost discriminative ability of AVID.

Extensive experiments and remarkable performance gains show the effectiveness of our method. Our method also achieves state-of-the-art transfer learning performance, both action recognition on UCF101 and HMDB51 datasets and sound recognition on ESC-50 and DCASE datasets, when pre-training on the subset of Audioset dataset.

2 Related Work

Instance Discrimination conducts contrastive learning by treating each sample as a separate class and pulling the positives (“similar”) closer while pushing the negatives (“dissimilar”) far away to learn instance-specific discriminative representation. MoCo [He et al., 2020] adopts an online encoder and a momentum encoder to receive two views of a sample as positive pairs. A queue-based dictionary is built to store negative samples. SimCLR [Chen et al., 2020a] pushes self-supervised pre-training to comparable effects as the supervised methods. They also carefully sort out the tricks for effective training, such as longer training time or stronger data augmentation. [Chen et al., 2020b] proposes MoCo-v2 and refreshes the performance of self-supervised learning once again. BYOL [Grill et al., 2020] adds a predictor to learn the map from the online encoder to the momentum encoder instead of displaying positive samples. Through the stop gradient mechanism, they skillfully discard the negatives. SimSiam [Chen and He, 2021] lets the target encoder and the online encoder be the same and points out that the predictor and the stop gradient mechanism are sufficient conditions for self-supervised training.

These works rely heavily on proper data augmentation setups. Instead, we focus on cross-modal instance discrimination, which avoids this issue through the inherent and pervasive synchronization between audio-visual messages. Furthermore, our goal is to be “semantics-aware”, not just instance-specific instance discrimination by introducing the idea of deep clustering [Zhan et al., 2020].

Self-Supervised Audio-Visual Representation Learning. Besides computer vision, several approaches attempt to learn audio-visual representations in a self-supervised manner. [Arandjelovic and Zisserman, 2017; Arandjelovic and Zisserman, 2018] propose to learn audio-visual representations by solving a binary verification problem that predicts whether the input audio-visual pair is matched or not. [Korbar et al., 2018; Owens and Efros, 2018] predict if audio-visual clips are synchronized in time, and [Morgado et al., 2020] predicts if the audio-visual clips extracted from the 360° videos are aligned in space. [Morgado et al., 2021a; Morgado et al., 2021b; Ma et al., 2021] improve upon the binary verification task, where the size of the contrastive set is decoupled from the mini-batch size, allowing it to be large. In this paper, we explore an online mining strategy, rather than random sampling, to construct a reliable contrastive set.

3 Audio-Visual Instance Discrimination

Let \( (A, V) \) be a collection of \( N \) videos, where \( A = \{a_i\}_{i=1}^N \), \( V = \{v_i\}_{i=1}^N \), each audio-visual pair \( \{a_i, v_i\} \) is from the same block of a video. Here, \( a_i \in \mathbb{R}^{H_a \times W_a} \) refers to the spectrogram of the raw audio waveform, where \( H_a \) is time step and \( W_a \) is frequency band. \( v_i \in \mathbb{R}^{F \times 3 \times H_v \times W_v} \) is the visual clip containing \( F \) RGB frames with a height of \( H_v \) and a width of \( W_v \). The learning objective of AVID is to encourage the representations of audio-visual clips to be similar if they come from the same temporal block of a video and vice-versa.
the same structure but independent parameters. Formally, the
parameters \( \delta \), gradients, and \( h(\cdot) \) represents the key encoder with learnable parameters \( \delta \), updated by the momentum. The superscript is used to distinguish various data streams from audio or visual modality, where the query encoder and key encoder from the same modality have the same structure but independent parameters. Formally, the formulation can be expressed as: \( q_i^v = f^v(v_i; \theta^v), q_i^a = f^a(a_i; \theta^a), k_i^v = h^v(v_i; \delta^v) \) and \( k_i^a = h^a(a_i; \delta^a) \).

Formally, AVID employs a Noise-Contrastive Estimation (NCE) [Gutmann and Hyvärinen, 2010] to learn two encoders independently, \( f(\cdot; \theta) \) and \( h(\cdot; \delta) \), by matching \( q_i \) with \( k_i \) against its contrastive set \( C_i = \{k_j \}_{j=1}^K \):

\[
\mathcal{L}_{\text{nce}}(q_i; k_i, C_i) = -\log \frac{\exp[\cos(q_i^v, k_i^v)/\tau]}{\sum_k \exp[\cos(q_i^v, k_i^v)/\tau]},
\]

where \( q_i = f(\cdot) \) and \( k_i = h(\cdot) \) are the outputs of query encoder \( f \) and key encoder \( h \) with parameters \( \theta \) and \( \delta \) respectively. \( \mathbf{\cos}(\cdot, \cdot) \) is the cosine similarity between a pair of representations, \( \tau \) is the temperature to control the concentration degree of distribution and \( K \) is the size of the contrastive set.

Minimising NCE loss means maximising a lower bound on the MI between audio-visual pairs, which allows the model to learn to discriminate whether audio-visual pairs are “similar” or “dissimilar”. The final objective \( \mathcal{L}_{\text{nce}} \) of AVID consists of two terms:

\[
\mathcal{L}_{\text{nce}} = \mathcal{L}_{\text{nce}}(q_i^v | k_i^v, C_i^v) + \mathcal{L}_{\text{nce}}(q_i^a | k_i^a, C_i^a),
\]

where the former indicates visual-to-audio component and the latter refers to audio-to-visual component.

There are two most recently proposed approaches for performing AVID, i.e., Cross-modal ID [Morgado et al., 2021a; Morgado et al., 2021b] and Multi-modal MoCo [Ma et al., 2021]. The diagram of these two approaches is depicted in Fig.2. We make a brief summary of the main differences and connections between the two.

**Momentum Update.** As shown in Fig.2, both Cross-modal ID and Multi-modal MoCo employ a momentum update to maintain consistency during training:

\[
u_{old} \leftarrow m u_{old} + (1 - m) u_{new},
\]

where \( m \in [0, 1) \) is a momentum coefficient, \( u \) is the content to be updated. The former utilizes a momentum-based moving average memory bank [Wu et al., 2018], while the latter exploits a momentum-based slowly-changing encoder.

Besides, Multi-modal MoCo performs a cross-modal online dictionary look-up, i.e., the encoded representations of the current mini-batch are enqueued and the oldest ones are dequeued.

Given a query, its contrastive set \( C_i \) in Eq.1 is constructed by randomly sampling from a semantically indistinguishable memory bank or queue-based dictionary as there are no annotations available. Such a random sampling strategy makes the contrastive set contain massive faulty negatives and undermines the primary goal of AVID, resulting in no improvement or even impairment when simply increasing the size of the contrastive set.

## 4 Active Contrastive Set Mining

We believe that the random sampling without semantic guidance is the main cause of the above problem when constructing the contrastive set. For this purpose, we introduce a novel semantic-sensitive library\(^1\) for mining the contrastive set with informative and diverse negatives for robust AVID. Such libraries serve three functions:

- playing the role of conventional memory bank or dictionary to store a series of representations on-the-fly;
- imposing the underlying semantic structures on the audio and visual samples;
- providing the semantic guidance when constructing the contrastive set.

To be concrete, different from a single library with semantic mixtures, we manage \( C \) independent libraries \( M = \{M_m\}_{m=1}^C \), where each library \( M_m \) corresponding to one semantics. They are constructed by a learnable classifier that predicts the pseudo-label of a query. For a query \( q_i \) with pseudo-label \( y_i \), we can construct its contrastive set \( C_i \) in a manner similar to supervised contrastive learning [Khosla et al., 2020]:

\[
C_i = \{k_j \mid k_j \in M_m, \forall m \neq y_i\},
\]

where \( j \in [1, K] \) and \( m, y_i \in [1, C] \). After every backward pass, its key \( k_i \) is used to update the corresponding \( M_{y_i} \) in a queue-based or momentum-based manner.

The samples with various pseudo-labels are put in separate libraries. As a result, one library can be viewed as one semantic “cluster”, where the samples in same library can be taken as anchors to describe this semantics. These semantic libraries can naturally serve as the semantic decision boundaries based on sample-anchor similarity:

\[
g_{i,j} = \sum_{m \in M_j} \frac{\exp[\cos(q_i, m)/\tau]}{\sum_{m=1}^C \sum_{m \in M_m} \exp[\cos(q_i, m)/\tau]},
\]

Given a query \( q_i \), its similarity-based semantics assignments are represented as \( g_i = [g_{i1}, \cdots, g_{ij}, \cdots, g_{iC}] \), where \( g_i \in \mathbb{R}^C \). With such potential memberships, we can train the classifier which aims to minimise \( g \), and the semantics predictions \( p_i \), where \( p_i \) is yielded by the classifier \( r(\cdot; \gamma) \) with the

\(^1\)We use library to collectively denote memory bank and dictionary. That is to say, the library can be updated in a queue-based or momentum-based manner.
learnable parameters $\gamma$, i.e., $p_i = r(q_i)$. We propagate the gradient back to the classifier only to avoid feature learning from unreliable semantics. With the updated classifier $r(\cdot)$, we update the predicted semantics $y_i$ in a maximum likelihood manner for the contrastive set construction in Eq. 4. At the same time, the samples are assigned to the library with the most similar anchors and each library holds its own attribute that makes it different from the others.

**Discussion.** Intuitively, at the beginning of training, the weak classifier provides near-random semantics predictions, so that our semantic libraries contain no semantic structures. At this point, the contrastive set constructed by Eq.4 approximates to a random sampling. With the progressive classifier, our semantic libraries cover increasingly clear semantic structures and provide effective semantic guidance when constructing the contrastive set, allowing our contrastive set to contain more informative and diverse negatives.

Furthermore, both Cross-modal ID [Morgado et al., 2021a; Morgado et al., 2021b] and Multi-modal MoCo [Ma et al., 2021] only optimise instance-specific discriminative representations, leading them to be not semantics sensitive and unaware of any potential non-linear intra-semantics variation. Compared with them, our ACSM maximises not only the instance-level diversity within the library, but also semantic-level compactness among libraries. The samples belonging to the same semantic library may share a common contrastive set. It means that the samples within the library are indirectly pushed closer in the representation space, resulting in more compact within-library representations.

**Hard Sample Mining.** The samples that are easily discriminated by the current model tend to contribute less to instance discrimination [Sheng et al., 2020]. Hard sample mining is important in the instance discrimination field but is especially essential in the audio-visual field. From the information-theoretic perspective [Chen et al., 2018], audio-visual messages in videos contain higher MI due to the higher dimensionality (i.e., temporal and multi-modal). The instances with higher $MI$ mean that they are easier to discriminate. For this purpose, [Bruno et al., 2018] perform the hard sample mining based on one key assumption: the smaller the time gap is between audio-visual clips of the same video, the harder it is to discriminate them. Instead of making such a assumption, we mine hard samples by emphasising the samples with ambiguous semantics. Given a query $q_i$ with pseudo-label $y_i^e$ at $e$th epoch, its semantic ambiguity is denoted as:

$$s_i^e = s_i^{e-1} + 1, \text{ if } y_i^e \neq y_i^{e-1} \quad (6)$$

Those samples that are frequently swapped across semantics will be assigned higher weights in Eq.2 to provide more useful discriminative clues. Overall, we apply our ACSM to Cross-modal ID [Ma et al., 2021] and Multi-modal MoCo [Morgado et al., 2021a; Morgado et al., 2021b].

## 5 Experiments

For quantitative evaluation, we evaluate both the pre-trained visual and audio representations using transfer learning. Concretely, these pre-trained representations are evaluated by linear probing, i.e., keep the pre-trained network fixed and train linear classifiers.

### 5.1 Experimental Setups

**Data preprocessing.** Video clips are extracted at a frame rate of 16fps with some standard data augmentation strategies, i.e., random multi-scale cropping, random horizontal flipping, color jittering and temporal jittering.

For audio clips, each audio with a duration of 2s is randomly sampled within 0.5s at a 24kHz sampling rate. Then, we convert the spectrogram of the audio track to a log scale,

| Methods | Pretraining Dataset | Architecture | UCF-101 | HMDB-51 |
|---------|---------------------|--------------|---------|---------|
| Shuffle&Learn [Misra et al., 2016] | UCF-101 | CaffeNet | 50.2 | 18.1 |
| OPN [Lee et al., 2017] | | VGG | 58.6 | 25.0 |
| ST-Order [Buchler et al., 2018] | | 3D-ResNet18 | 59.1 | 26.7 |
| CMC [Tian et al., 2020] | Kinetics-600 | 16 × 122² | 79.5 | 44.6 |
| CBT [Sun et al., 2019] | | 3D-ResNet18 | 62.9 | 33.7 |
| 3D-RotNet [Jing et al., 2018] | | R(2+1)D-18 | 72.4 | 30.9 |
| ClipOrder [Xu et al., 2019] | | 3D-ResNet34 | 75.7 | 35.7 |
| DPC [Han et al., 2019] | Kinetics-400 | 16 × 224² | 74.4 | 47.8 |
| $L^3$-Net [Arandjelovic and Zisserman, 2017] | | 2D-Conv | 85.8 | 56.9 |
| AVIS [Korbar et al., 2018] | | MC3 | 74.2 | 39.0 |
| xID [Morgado et al., 2021b] | | R(2+1)D-18 | 82.3 | 49.1 |
| Robust-xID [Morgado et al., 2021a] | | 8 × 224² | 81.9 | 49.5 |
| xID + CMA [Morgado et al., 2021b] | | 8 × 224² | 83.4 | 50.2 |
| MnMoCo [Ma et al., 2021] | | 8 × 224² | 83.5 | 49.4 |
| MnMoCo + AS [Ma et al., 2021] | | 8 × 224² | 85.6 | 58.6 |
| xID + ACSM (Ours) | | 8 × 224² | 86.7 | 60.4 |
| MnMoCo + ACSM (Ours) | | 8 × 224² | 88.4 | 61.3 |
SoundNet [Aytar et al., 2016] & ConvNet-8 & 74.2 & 88.0 & 79.3 & 93.0
$L^3$-Net [Arandjelovic and Zisserman, 2017] & SoundNet & ConvNet-9 & 82.0 & 92.4 & 78.8 & 93.6

Table 3: Action recognition accuracy of linear probing on Kinetics and sound recognition accuracy of linear probing on ESC-50, where the evaluated representations come from various blocks of the encoder, ↑ indicates the increase in accuracy.

| Method | Kinetics-400 | ESC-50 |
|--------|--------------|--------|
|        | block1 | block2 | block3 | block4 | block1 | block2 | block3 | block4 |
| xID    | 19.80 | 26.98 | 34.81 | 39.95 | 67.25 | 73.15 | 74.80 | 75.05 |
| xID + ACSM | 1.26↑ | 1.69↑ | 2.37↑ | 3.56↑ | 3.85↑ | 5.56↑ | 7.12↑ | 8.71↑ |
| MmMoCo | 20.12 | 26.48 | 35.35 | 39.75 | 66.86 | 72.69 | 74.93 | 76.56 |
| MmMoCo + ACSM | 0.94↑ | 1.76↑ | 2.68↑ | 3.35↑ | 4.11↑ | 5.98↑ | 6.94↑ | 8.58↑ |

5.2 Comparisons with SOTAs

It is impractical to give a direct comparison to all approaches due to the enormous differences in the experimental setups, e.g., using diverse encoder architectures, pre-training on various datasets and fine-tuning with different resolutions. To empower solid comparisons, we list these differences for each method in Tab.1 and Tab.2, where Cross-modal ID and Multimodal MoCo are abbreviated as xID and MmMoCo, respectively. The methods with the best and second-best performance are indicated in boldface.

Visual representations. To evaluate the visual representations, we compare the transfer performance of action recognition with previous self-supervised methods on UCF-101 and HMDB-51 datasets. We follow [Morgado et al., 2021b] and report top-1 accuracy of video-level predictions by averaging the clip-level accuracy of 10 uniformly sampled clips per test sample in Tab.1.

Audio representations. We evaluate the audio representations on ESC-50 and DCASE datasets for sound recognition. A linear one-vs-all SVM classifier is trained on the audio representations obtained by the pre-trained models at the final layer before pooling. We report top-1 accuracy of sample-level predictions by averaging 10 clip-level predictions in Tab.2.

As we can see, our approach outperforms other SOTA approaches. Compared with MmMoCo + AS [Ma et al., 2021],
Figure 3: Effects of learning with different size of contrastive set $K$.

Figure 4: Effects of learning with different number of semantic libraries $C$. Both Cross-modal ID and Multi-modal MoCo are used as baselines.

the current best performing method on self-supervised audio-visual representation learning, our model outperforms it by 2.8% on UCF101, 2.7% on HMDB51, 2.9% on ESC-50 and 0.9% on DCASE. Compared with xID [Morgado et al., 2021b] and MmMoCo [Ma et al., 2021] (12th and 15th rows in Tab.2, 8th and 11th rows in Tab.1), our ACSM achieves impressive performance improvements on both action recognition and sound recognition.

5.3 Ablation Study

Random sampling vs. Active mining. We evaluate the learned visual and audio representations in terms of the various blocks of the video and audio encoder. Separately, we employ the Kinetics-400 for action recognition and ESC-50 for sound recognition. Tab.3 reports the top-1 accuracy by averaging predictions over 25 clips per video. Introducing our ACSM in Cross-modal ID and Multi-modal MoCo, the accuracy of both action recognition and sound recognition has been improved to varying degrees. The increase in recognition performance validates the benefits of a semantically sensitive library over a semantically indistinguishable one, resulting in better generalization of the learned representations to downstream tasks.

Hard Sample Mining. We evaluate the learned visual and audio representations according to different hard sample mining strategies. We employ the representations from the block4 of the encoder to evaluate the performance of transfer learning on Kinetics-400 and ESC-50 datasets. [Bruno et al., 2018] execute hard sample mining based on the time gap, i.e., $1 \sim 200$ epochs with easy samples only and $201 \sim 400$ epochs with $25\%$ hard samples and $75\%$ easy samples. To make fairer comparisons, we only execute our hard sample mining defined by Eq.6 in $201 \sim 400$ epochs. The top-1 accuracy by averaging predictions over 25 clips per video is stated in Tab.4. As expected, hard sample mining improves the quality of learned representations. Compared with Bruno et al., 2018, our mining strategy based on semantic ambiguity is more effective.

Effects of the size of the contrastive set. Both memory bank and queue-based dictionary decouple its size from the mini-batch size, allowing it to be large. Increasing the size of the contrastive set has been shown to improve the learned representations [McAllester and Stratos, 2020]. We empirically study the effects of learned representations with different $K$ on ESC-50. The results are shown in Fig.3. As the size of the contrastive set increases, performance almost always improves to varying degrees. However, the performance gains of MmMoCo gradually saturate. The performance of xID even shows a small drop when $K$ grows from 4096 to 8192. By contrast, our ACSM samples the contrastive set under the guidance of the semantic library, allowing the learned representations to maintain a steady growth in accuracy.

Effects of the number of the semantic library. There is no universal principle to help determine the proper semantic library number on auxiliary tasks so to maximise their benefits. We empirically investigate the effects of learned representations with different $C$ by using cross-modal ID and multi-modal MoCo as baselines. The results are demonstrated in Fig.4. All models with our semantic library can learn better representations, resulting in higher performance on downstream tasks. This again suggests the defects of randomly sampled contrastive set on high-level semantic understanding. In addition, we also observe that the accuracy on ESC-50 grows slower and slower with the increase of semantic library number.

6 Conclusion

To alleviate the false negatives caused by random sampling in AVID, we propose an active contrastive set mining (ACSM) that aims to learn a semantics-aware, not just instance-specific discriminative space. Also, we introduce a novel hard sample mining strategy based on semantic ambiguity. Our method can be elegantly extended to two most recent state-of-the-art AVID models and achieves impressive performances.
