CONTRASTIVE SELF-SUPERVISED LEARNING OF GLOBAL-LOCAL AUDIO-VISUAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive self-supervised learning has delivered impressive results in many audio-visual recognition tasks. However, existing approaches optimize for learning either *global* representations useful for high-level understanding tasks such as classification, or *local* representations useful for tasks such as audio-visual source localization and separation. While they produce satisfactory results in their intended downstream scenarios, they often fail to generalize to tasks that they were not originally designed for. In this work, we propose a versatile self-supervised approach to learn audio-visual representations that can generalize to both the tasks which require global semantic information (e.g. classification) and the tasks that require fine-grained spatio-temporal information (e.g. localization). We achieve this by optimizing two cross-modal contrastive objectives that together encourage our model to learn discriminative global-local visual information given audio information. To show that our approach learns generalizable video representations, we evaluate it on various downstream scenarios including action/sound classification, lip reading, deepfake detection, and sound source localization.

1 INTRODUCTION

Self-supervised learning aims to learn representations of data that generalize to a large variety of downstream tasks. Recently, contrastive self-supervised learning (CSL) has achieved impressive results on several computer vision tasks (Oord et al., 2018; Hjelm et al., 2018; He et al., 2020; Chen et al., 2020). In CSL, the choice of "views" determines the types of information that the representation captures (Bachman et al., 2019), as the framework learns representations that focus on the shared information between views. It has been demonstrated that the optimal choice of views depends critically on the downstream task (Tian et al., 2020). Therefore, existing works mainly focus on finding different views tailored for the intended downstream tasks. For example, when tailoring views for action classification, Hjelm & Bachman (2020) extends DIM (Hjelm et al., 2018) to the spatio-temporal setting by assuming that global and local information useful for action classification (i.e, global semantics) should be invariant across time and space in a single sample.It thus takes both global and local features extracted from the same video sample as positive pairs. In the multimodal setting, several approaches utilize audio-visual correspondence from videos (Morgado et al., 2020). Such CSL approach is based on an assumption that information needed for audio/video classification should be shared between the two modalities.

Although they achieve impressive results in their intended downstream tasks, existing approaches often fail to generalize to tasks that they were originally designed for. For example, in lip reading (Chung & Zisserman, 2016), the desired information is the fine-grained spatio-temporal representation of the human face. However, if we directly apply existing CSL approaches, the shared information across views ("global semantics") is a "face", while the useful information, i.e., the "lip movements", will be bottlenecked as they are, by the nature of the task, changing across views.

Motivated by this, we propose a versatile CSL approach to learn representations that can generalize to both scenarios that require global representations (e.g., classification) and scenarios that require local representations (e.g., localization) (see Fig. 1). Our approach, which we call *global-local cross-modal* (GLCM) contrastive learning, has four key properties that we assume to be important for our learning objective: 1) observations from the same time span of a video should reflect the same content regardless of modalities; 2) the same observations captured at different time scales can

reflect both global and local information; 3) when learning on a local temporal scale, the contrasting views should only share the time-varying information (e.g. only the moving lip) while ignoring globally invariant information; 4) multi-scale (global-local) observations can be trained jointly in a collaborative way so that the captured information in either scale can be reused.

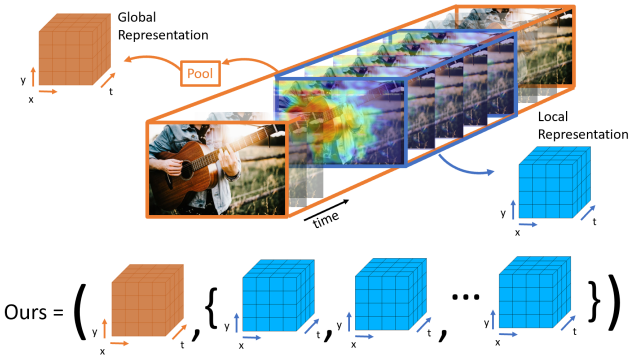

Figure 1: While many self-supervised approaches optimize for high-level *or* low-level tasks, we present an approach to learn both global *and* local representations from video.

We formulate our GLCM objective using two cross-modal contrastive losses in a multi-scale setting. Specifically, we generate global and local views of a visual sequence using different sampling rates. The audio sequence is used as an anchor to contrast with global and local visual features, respectively. To meet the first property, the losses are computed at the same temporal scale, i.e. $z_a^g \leftrightarrow z_v^g$ and $z_a^l \leftrightarrow z_v^l$, such that, given the same video sequence, the learned z_v^g and z_v^l can reflect both global and local information; this latter part achieves the second property. To achieve the third property, the local contrastive loss ($z_a^l \leftrightarrow z_v^l$) is computed by considering only the audio-visual features that lie in the same time window as positive pairs; the others are all negative pairs. Finally, we utilize information captured at the global scale (e.g. localized sound source) to assist efficient learning at the local scale.

We show that the pretrained GLCM can capture both global and fine-grained spatio-temporal information from audio-visual signals, and that enables easy adaptation to a wide variety of downstream tasks. We evaluate our proposed approach on tasks that needs local spatio-temporal information, i.e lip reading, deep-fake detection, and sound-source localization, and also discriminative tasks that needs global information, i.e. action classification and audio-event classification.

2 Related Work

The goal of self-supervised representation learning is to create data transformations that are effective for solving downstream tasks (Doersch & Zisserman, 2017; Kolesnikov et al., 2019). In audio-visual understanding this involves "global" task such as recognizing scenes Owens et al. (2016) and "local" tasks such as detecting the source of audio within a clip (Senocak et al., 2018) and separating this from other sources (Ephrat et al., 2018). Self-supervised learning has enabled strong performance on many tasks and in cases produced comparable results to supervised learning (Chen et al., 2020; Caron et al., 2020). Contrastive learning is one of the most effective approaches to self-supervised learning to date and many effective methods leverage multiple views of the same data (Tian et al., 2019; Hjelm & Bachman, 2020). Views could correspond to multiple perspectives within the same modality (e.g., different frames of a video) (He et al., 2020; Hjelm & Bachman, 2020; Han et al., 2019) or perspectives from differnet modalities (e.g., depth and RGB images, visual and textual signals) (Tian et al., 2019; Sun et al., 2019). In videos, audio and visual data are both complementary and considering their correspondence (Owens et al., 2016), could be considered to be different views of the same data. Thus there exists a collection of works that leverages the natural correspondence of audio and visual signals to perform contastive learning (Asano et al., 2020; Korbar et al., 2018; Alwassel et al., 2019; Morgado et al., 2020; Patrick et al., 2020). Most of the self-supervised contrastive learning approaches aim to capture high-level information from observations. It has been empirically demonstrated that such learned information is very effective for "discrimination tasks" (classification). However, in tasks that needs local information the learned representations may not perform well. Chen et al. (2020) and Hjelm et al. (2018) propose to leverage local information to perform contrastive learning, and validate that learning approach such as this can further improve the performance on image classification tasks. DIM (Hjelm et al., 2018) has been extended to multi-scale (Bachman et al., 2019) and video data Hjelm & Bachman (2020).

However, evaluation was still limited to "discriminative" tasks (image classification and video event classification). There is little evidence in the self-supervised contrastive literature that these models will adapt well to the local information and more work is needed.

3 APPROACH

We propose using the audio and visual channels as cross-modal views of video data. As we aim to learn both local and global temporal information, we also utilize the same visual sequence processed at different sampling rates to reflect the same observation at different temporal scales. Given that we want each signal to capture complementary views of the same data, we use different encoders to extract the representations from the audio sequence (E_a), subsampled visual sequence (E_v^g) and full sampling rate visual sequence (E_v^l). The question is, then, how to design a contrastive loss to learn representations from these different views. We achieve this goal by joint training the model using two contrastive losses, a global contrastive loss and a local contrastive loss. As shown in Fig. 2, the global loss is computed by contrasting audio signals with subsampled visual sequence (Sec.3.1); while the local loss is computed by contrasting audio signals with visual sequence at a full sampling rate (Sec.3.2). To satisfy property 4 (i.e., global-local pathway can be trained jointly in a collaborative way), we propose a spatially-aware attention pooling mechanism to effectively reuse the information that was captured from the global pathway to the local pathway (Sec.3.3).

3.1 GLOBAL CONTRASTIVE OBJECTIVE

We design the global contrastive objective to capture slowly changing information with high audio-visual correlation. We use video sequences captured at a low sampling rate, which will inevitably lack local temporal information. E_a encodes an audio sequence into the audio embedding $z_a \in R^{T \times F}$, where F is the frequency, and T is the sequence length. After temporal global pooling, it becomes $z_a \in R^{1 \times F}$. Similarly, we perform global temporal pooling on features encoded by the global visual encoder E_v^g, which produces the global visual embedding $z_v^g \in R^{1 \times H \times W \times C}$. Note that, for the visual feature, we perform global pooling only along the temporal dimension while keeping the spatial dimension intact. The reason is that when learning from a global temporal space, we can use the model capacity to capture more information from the spatial dimension. To compute the global contrastive loss, we consider the audio feature z_a and the visual feature z_v^g that come from the same video sample as positive pairs, while the ones come from different video samples are negative pairs. In order to encourage the model to also capture spatial information, we adopt MIL-NCE (Miech et al., 2020) to compute the loss. Specifically, we consider all $H \times W$ spatial grids in

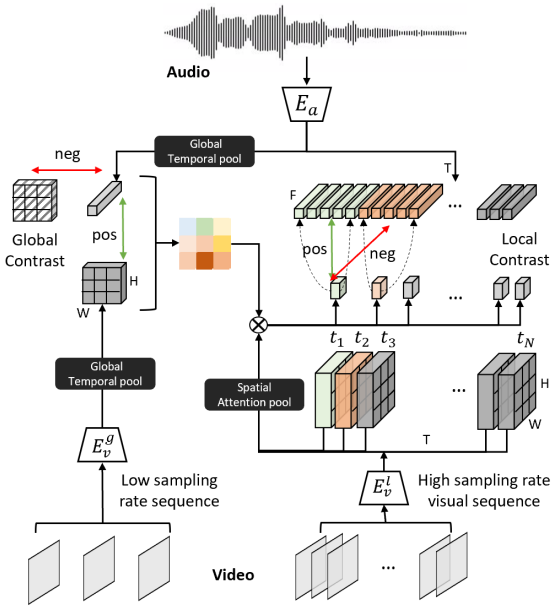

Figure 2: Our global-local contrastive learning architecture. For clarity we omit the "channel" dimension, and only show the spatial and temporal dimensions. In the "global contrast" pathway, the visual features are shown with different shading patterns (e.g., diagonal-hatch, filled gray) and indicate those that come from different video samples. In the "local contrast" pathway, we use colors to indicate different time windows. For ease of illustration, we only show the non-synchronized negative pairs in the local contrast pathway. In our implementation we also incorporate easy negative pairs, i.e. features that come from different samples.

z_v^g as the instances, and therefore, instead of just taking a single audio-visual positive pair $z_a \leftrightarrow z_v^g$, the new positive pair becomes multiple visual instances $z_v^g[i]_{H\times W}$, i.e. $z_a \leftrightarrow z_v^g[i]_{H\times W}$. The loss is then defined as:

$$L_g = -log(\frac{\sum_{z_v^g \in \mathcal{P}} exp(z_a^T z_v^g)}{\sum_{z_v^g \in \mathcal{P}} exp(z_a^T z_v^g) + \sum_{z' \in \mathcal{N}} exp(z_a'^T z_v'^g)}) \tag{1}$$

where $\mathcal{N}$ is a set of negative audio and visual pairs, $\mathcal{P}$ is a set of spatial grids in z_v^g.

3.2 Local Contrastive Objective

We design the local contrastive objective to capture fine-grained spatio-temporal information that are sensitive to temporal changes while being invariant to different modalities. We thus contrast between local audio feature z_a and local spatio-temporal visual feature z_v^l. Specifically, we consider the temporal local audio and visual features that lie in the same time window to be the positive pairs, $z_a[t] \leftrightarrow z_v^l[t]$, where t represents features in the time window t. As shown in Figure 2, the video and audio features shaded in the same color refers to that in the same time window. The features in different time window (e.g. green and orange blocks) are considered as negative pairs even if they are from the same video sample. In this way, we let the modalities to control the shared information between them to be just the time-changing part. We obtain the local audio feature by using the same audio encoder E_a but without global temporal pooling. The local visual feature z_v^{local} is obtained by feeding the visual sequence with a high sampling rate into the local visual encoder E_v^l, which produces the visual feature $z_v^l \in R^{T\times H\times W\times C}$. We then perform pooling just along the spatial dimension while keeping temporal scale the same, the visual feature becomes $z_v^l \in R^{T\times 1\times 1\times C}$.

As audio signals naturally have a higher sampling rate than visual signals, visual feature at a single time step will be mapped to multiple audio feature slices. As shown in Figure 2, at time t_1, the visual feature $z_v^l[t_1]$ (green block) corresponds to multiple audio features $z_a[t_1]_{t\in M}$ (green blocks), where $M = 5$ in Figure 2. Because of the lack of strict one-to-one mapping, here we also utilize MIL-NCE (Miech et al., 2020) to compute the contrastive loss so that it holds the robustness in temporal mapping. Specifically, we use sliding window of size M to map each time step of visual feature to a window of audio feature slices. Then the positive pair is considered as a visual feature and the corresponding window of audio feature slices. The goal of using MIL-NCE here is different from using it the global contrastive loss. In the global contrastive loss, we aim to let the network capture spatial information. While in the local contrastive objective, the goal of using MIL-NCE is to mitigate the missing strict temporal mapping problem. The loss is therefore defined as:

$$L_l = -log(\frac{\sum_{z_a \in \mathcal{Q}} exp(z_a^T z_v^l)}{\sum_{z_a \in \mathcal{Q}} exp(z_a^T z_v^l) + \sum_{z' \in \mathcal{N}} exp(z_a'^T z_v'^l)}) \tag{2}$$

where $\mathcal{Q}$ is a set of audio features slices in the same time window with z_v^l, $\mathcal{N}$ is a set of negative audio and visual pairs.

3.3 Spatially-aware attention pooling

To satisfy property 4 (Sec. 1), we want the global and local pathways to be jointly trained in a collaborative way. As discussed in Section 3.1, when computing the global contrastive loss, we focus on the spatial dimension. Therefore, we can utilize spatial information captured from the global pathway to assist the local contrastive loss. Specifically, we take the correlation score (i.e. dot product between the audio feature and each of the visual features in a spatial grid) computed in the global pathway as the attention score; intuitively, this reflects the regions where the sound come from. For example, in a "human speech" video, the lip region will have a relatively higher score than the background, and in a "playing guitar" video, fingers on the guitar will have high scores. We thus use the $H \times W$ score as the attention map (as shown in Figure 2). We utilize this attention map to perform a spatial attention pooling on the local visual features at each time step $P_{atten}(z_v^l[t])$. Comparing with regular spatial average pooling, it helps the network pay more attention to parts within a frame with high audio-visual correspondence. This way, the efficiency of the local contrastive loss can be much improved. We empirically demonstrate the effectiveness of the spatial attention pooling mechanism (Table 1 and Table 2).

4 EXPERIMENTS

Implementation details. We use 3D-ResNet18 (Hara et al., 2018) as our visual encoders (E_v^g and E_v^l) and 1D-ResNet18 as an audio encoder. We employ Batch Normalization (BN) (Ioffe & Szegedy, 2015) in all our encoders. All models are trained end-to-end with the ADAM optimizer (Kingma & Ba, 2014) with an initial learning rate $\gamma = 10^{-3}$ after a warm-up period of 500 iterations. We use 16 NVIDIA Tesla P100 GPUs with a batch size of 32 for our experiments. For pretraining, we preprocess video frames by sampling at 10 FPS and applying random cropping, horizontal flipping, gray-scaling, and temporal jittering. We resize video frames to 3-channel images of 112×112; we set the clip length to 32 frames for the local visual pathway, and $\frac{1}{8}$ sampling rate for the global path way (8 frames). For the audio channel, we extract mel-spectrograms from the raw waveform using the LibROSA library and get a $80 \times T$ matrix with 80 frequency bands; T is proportionate to the length of an audio clip. We then segment the mel-spectrogram according to the corresponding video clips to ensure temporal synchrony. We treat the mel-spectrograms as an 80-channel 1D signal. To compute the global MIL-NCE loss, we use the features at 16×16 spatial dimensions. To compute the local contrastive loss, we adopt a temporal window of size three without overlap.

Datasets. We use three datasets for pretraining. For speech related tasks, we pretrain our model on a combination of the "pretrain set" of LRS2 Chung et al. (2017) and LRS3 (Afouras et al., 2018) datasets. For deepfake detection, we pretrain our model on a subset of 36K clips on the DFDC (Dolhansky et al., 2019) dataset. For classification tasks, we pretrain our model on a subset of 240K clips from Kinetics-700 (Carreira et al., 2019). As for downstream tasks, we evaluate our models on action recognition using UCF101 (Soomro et al., 2012) and HMDB51 (Kuehne et al., 2011), and on sound classification using ESC50 (Piczak, 2015b). For lip reading, we evaualte our model on both LRW (Chung & Zisserman, 2016) and LRS2 Chung et al. (2017). For deepfake detection, we evaluate our model on another subset of DFDC which has no overlap with our pretraining dataset.

4.1 DOWNSTREAM SCENARIO: LIP READING

Lip reading, i.e., visually recognizing the speaker's utterance, is a challenging task because lip movements for different letters can be visually similar to each other (e.g., b and p, d and t). This requires the learned visual representation to contain fine-grained spatio-temporal information, rather than global semantics. In evaluating our pretrained model on the lip reading task, we focus on the following: 1) Does our approach successfully learn fine-trained spatio-temporal visual information? 2) Should we use either the local or global representation alone, or combine them? 3) Can the local spatial information captured from the global-pathway help learn local spatio-temporal features? We will start by first showing our experiment details, and then analyze and discuss the results.

Data preprocessing: For a fair comparison with state-of-the-art (SOTA) approaches, we use the same data processing protocol of Zhang et al. (2019). For LRW and LRS2, we detect 68 facial landmarks in each video frame using the dlib library (Castelli & Pagano, 2002). We align detected faces by affine transforming frames based on two landmarks outside the eyes and the lowest landmark on the nose. Finally, a lip-centered image of size 112×112 is cropped from the aligned face. The width of the lip is normalized to occupy $\frac{1}{3}$ of the image width. During finetuning, we apply random horizontal flipping.

Comparison with SOTA on LRW and LRS2. We answer the first question by comparing our approach with SOTA lip reading methods. For LRW, we evaluate on a 500-way word classification task and report top-1 accuracy. For LRS, we report the word error rate (WER). Table 1 shows that our approach outperforms SOTA approaches on both LRS and LRW by a large margin, i.e. 3.1% WER reduction on LRS and 3.07% accuracy improvements on LRW. It validates that our proposed training approach can effectively capture the fine-grained spatio-temporal information necessary for lip reading.

Comparison of global and local representations. To demonstrate the importance of learning local information during pretraining, we evaluate a baseline that uses only the global visual feature; see "Ours global" in Table 1. The results show that when we discard local information, the performance significantly drops on both LRS and LRW. This validates our claim that optimizing only for global representation during pretraining generalizes poorly to the tasks that require local information.

The role of spatial attention. In our pretraining pipeline, we can leverage the global contrastive objective to capture local spatial information and use it as an attention map to assist local representation learning. Intuitively, our attention maps measure the amount of audio-visual correlation; such attention maps can highlight discriminative face regions useful for lip reading. We demonstrate the quality of our attention maps by replacing lip bounding boxes typically used in lip reading with our attention map. Specifically, instead of extracting features from the cropped lip region, we extract features from the entire frame (no lip cropping) and use our attention map to pool the features spatially. Note that the purpose of this experiment is to evaluate the quality of attention maps; we use audio signal just to obtain attention maps and discard it for word classification. The results ("Ours Attention") show that it achieves results comparable to our best setting. On LRW, it even outperforms SOTA approaches without relying on lip region detectors. It indicates that using global and local information in a collaborative way can yield good performance. In the lip reading task, local spatial information makes the local-contrast pathway pay more attention to lip movement, and thus achieves comparable effects of using a lip region detector.

Model	Modality	LRS (WER)	LRW (Acc.)
LRW (Chung & Zisserman, 2016)	RGB	N/A	61.10
Res.LSTM (Stafylakis & Tzimiropoulos, 2017)	RGB	N/A	83.00
TwoStream (Weng & Kitani, 2019)	RGB&OF	N/A	84.07
DFTN (Xiao et al., 2020)	Gray&DF	N/A	84.13
TM-sep2seq (Afouras et al., 2018)	RGB	49.8	N/A
TM-CTC (Afouras et al., 2018)	RGB	65	N/A
STF (Zhang et al., 2019)	RGB	51.7	83.7
WAS (Chung et al., 2017)	RGB	70.4	76.2
Ours global	RGB	58.3	71.8
Ours local	RGB	47.2	**87.2**
Ours global+local	RGB	**46.7**	86.9
Ours attention (no lip cropping)	RGB	51.2	85.7

Table 1: Lip reading comparison with SOTA. LRW (top-1 accuracy, the higher the better) LRS(WER, the lower the better). "RGB": RGB image, "Gray": gray scale image, "OF": optical flow, "DF": deformation flow. "global" and "local" refer to global and local feature. For "Ours Attention", we do not do lip cropping and instead perform spatial attention pooling.

4.2 Downstream Scenario: Deepfake detection

Deepfakes tend to be characterized by audio-visual inconsistencies such as a misalignment between lip motions and audio, unnatural facial and lip appearance/movements or assymmetry between facial regions such as the left and right eyes. Such artifacts can be detected through local spatio-temporal features. We thus use our pretrained model and finetune it on the DFDC dataset. We use video-wise Area Under the Curve (AUC) as our evaluation metric. We follow the same data preprocessing protocol of SOTA approaches in this task. We perform face detection to crop the face region in each video frame.

The results are shown in Table 2. For a fair comparison, we use the same training and test video lists as Chugh et al. (2020). Among all the compared approaches, Chugh et al. (2020) and Mittal et al. use both visual and audio sequences, while the other approaches use only the visual sequences for detection. As we can see, when using only visual signal, our approach outperforms all previous SOTA approaches (AUC=95.5). However, when we use both audio and visual signals, the performance drops from 95.5 to 94.6. We suspect that the fake videos have only had the visual modality manipulated, so adding audio signal does not add more useful information for discriminating between 'real' and 'fake' samples. In fact it may "distract" the network, and thus impair the performance. We also use the audio signal to obtain the attention map and perform a spatial attention pooling. We find that without face cropping, when we use spatial attention pooling, the performance is comparable to our best performance. More importantly, it outperforms SOTA approaches that rely on face cropping.

Model	Visual	Audio	AUC
Capsule (Nguyen et al., 2019b)	✓		53.3
Multi-task (Nguyen et al., 2019a)	✓		53.6
HeadPose (Yang et al., 2019)	✓		55.9
Two-stream (Zhou et al., 2017)	✓		61.4
VA-MLP (Matern et al., 2019)	✓		61.9
VA-LogReg	✓		66.2
Meso4 (Afchar et al., 2018)	✓		75.3
Xception-c40 (Rossler et al., 2019)	✓		69.7
Xception-c23	✓		72.2
FWA (Li & Lyu, 2018)	✓		72.7
DSP-FWA (Li & Lyu, 2018)	✓		75.5
Siamese (Mittal et al.)	✓	✓	84.4
MDS (Chugh et al., 2020)	✓	✓	91.5
Ours	✓		95.5
Ours Atten. (no face cropping)	✓	✓	95.0
Ours Multimodal	✓	✓	95.6

Table 2: Comparison with SOTA approaches on deepfake detection. For training on DFDC, we use the same training and testing list of Chugh et al. (2020). Mittal et al. use the same number of randomly sampled training data (180K). All the numbers of the other SOTA approaches are collected from Chugh et al. (2020).

4.3 Downstream Scenario: Sound Source Localization

To investigate how well local spatial information is captured through the global audio-visual contrast, we leverage the pretrained audio and global video encoders to perform sound source localization. The task of sound source localization is to locate objects making sound. To achieve this goal, the network should capture the audio-visual correlation in a spatio-temporal grid. We thus use the attention map obtained by our pretrained model to visualize the sounding source in each frame. To investigate further, we use the Kinetics-sound (Carreira et al., 2019) dataset as videos in the dataset generally have a high-level of audio-visual correspondence. Specifically, we add another softmax layer on the obtained attention map, and then do bilinear interpolation of the 16×16 attention map back to the original image size, i.e. 192×192.

Fig. 3 shows that our learned attention maps tend to localize sounding sources in videos accurately. Especially, when visual content is highly related to the corresponding audio signal, our model performs particularly well. For example, the first row (frames from "playing instruments" videos) shows that our model can precisely localize the sounding region. For the other activities, like "baby talking," "playing basketball," "running," our model highlights regions with humans. We find that the attention map incorrectly highlights regions on samples that have ambiguous audio-visual relation. We show failure cases in the last two frames of the third row. As we can see, the there is no visual content that clearly relates with the audio signal, and thus the model fails to find sounding sources.

4.4 Downstream Scenario: Action and Sound Classification

To evaluate the effectiveness of our approach in learning discriminative global spatio-temporal representations, we use our pretrained model on action and sound classification tasks.

For action classification, we compare with SOTA approaches pretrained on the same dataset only (Kinetics-700). We find that on both UCF101 and HMDB51, our approach achieves a new state-of-the-art, i.e., 91.2% on UCF101 and 62.1% on HMDB51. We also find that the global features yield good performance, while using the local features alone leads to inferior results. This result suggests that the information captured is different when using different learning method. When we perform 'global contrast', the network is encouraged to capture more abstract semantics to discriminate the same sample representation with others. When performing 'local contrast', the information captured is the temporally changing part, which may not be directly related to the 'classification' task. The

Figure 3: Visualization of attention maps showing results for the source source localization task.

same observation can also be made in sound classification. As we can see, the best performance also comes when we use the 'global' features. When we combine the 'local' features for both tasks, the performance improves, but only a small amount.

Method	UCF101	HMDB51	ECS50
Scratch	46.5	17.1	N/A
Scratch Random Forest (Piczak, 2015b)	N/A	N/A	44.3
Scratch ConvNet (Piczak, 2015a)	N/A	N/A	64.5
Scratch ConvRBM (Sailor et al., 2017)	N/A	N/A	86.5
Supervised	82.8	46.7	N/A
RotNet3D (Jing & Tian, 2018)	62.9	33.7	N/A
AVTS (Korbar et al., 2018)	85.8	56.9	76.7
MotionPred (Wang et al., 2019)	61.2	33.4	N/A
ST-Puzzle (Kim et al., 2019)	65.8	33.7	N/A
ClipOrder (Xu et al., 2019)	72.4	30.9	N/A
CBT (Sun et al., 2019)	79.5	44.6	N/A
DPC (Han et al., 2019)	75.7	35.7	N/A
XDC (Alwassel et al., 2019)	84.2	47.1	78.0
SeLaVi (Asano et al., 2020)	83.1	47.1	N/A
AVID (Morgado et al., 2020)	87.5	<u>60.8</u>	<u>79.1</u>
GDT (Patrick et al., 2020)	<u>89.3</u>	60.0	N/A
Ours global	91.0	**62.1**	79.5
Ours local	80.3	52.1	75.1
Ours global+local	**91.2**	61.9	**79.7**

Table 3: Comparison of SOTA approaches on action classification and sound classification. We specify pretraining dataset and the number of samples used if they are reported in the original papers (N/A: not available). Results in 'bold' indicate the best results and 'underline' indicate the second best results.

5 Conclusion

We have presented a contrastive self-supervised approach to learning global and local audio-visual representations. Using audio and low-sampled and high-sampled video sequences as separate "views" of the data, we find that the learned representations can generalize well to tasks that involve global semantic understanding and fine-grained spatio-temporal understanding. We perform experiments on lip reading, deepfake detection, sound source localization, and action/sound classification tasks and in each case achieve strong results.

REFERENCES

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7. IEEE, 2018.

T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. In *arXiv preprint arXiv:1809.00496*, 2018.

Humam Alwassel, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019.

Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *arXiv preprint arXiv:2006.13662*, 2020.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pp. 15535–15545, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

Donatella Castelli and Pasquale Pagano. Opendlib: A digital library service system. In Maristella Agosti and Costantino Thanos (eds.), *Research and Advanced Technology for Digital Libraries*, pp. 292–308, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. *arXiv preprint arXiv:2005.14405*, 2020.

J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3444–3453. IEEE, 2017.

Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2051–2060, 2017.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *ICCV*, 2019.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

R. Devon Hjelm and Philip Bachman. Representation learning with video deep infomax. *arXiv preprint arXiv:2007.13278*, 2020.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

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

Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *AAAI*, 2019.

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

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1920–1929, 2019.

Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, 2018.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011.

Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 83–92, 2019.

A. Miech, J. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9876–9886, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00990. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00990`.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues.

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. *arXiv preprint arXiv:2004.12943*, 2020.

Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019a.

Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2307–2311. IEEE, 2019b.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European conference on computer vision*, pp. 801–816. Springer, 2016.

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.

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

Karol J Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, 2015b.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–11, 2019.

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

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4358–4366, 2018.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*, 2017.

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

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *CVPR*, 2019.

Xinshuo Weng and Kris Kitani. Learning spatio-temporal features with two-stream deep 3d cnns for lipreading. *arXiv preprint arXiv:1905.02540*, 2019.

Jingyun Xiao, Shuang Yang, Yuanhang Zhang, Shiguang Shan, and Xilin Chen. Deformation flow based two-stream network for lip reading. *arXiv preprint arXiv:2003.05709*, 2020.

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, 2019.

Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8261–8265. IEEE, 2019.

Xingxuan Zhang, Feng Cheng, and Shilin Wang. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1831–1839. IEEE, 2017.