Contrastive Learning of Global-Local Video Representations

Shuang Ma  
Microsoft  
Redmond, WA, USA

Zhaoyang Zeng  
Sun Yat-sen University  
Guangzhou, China

Daniel McDuff  
Microsoft Research  
Redmond, WA, USA

Yale Song  
Microsoft Research  
Redmond, WA, USA

Abstract

Contrastive learning has delivered impressive results for various tasks in the self-supervised regime. However, existing approaches optimize for learning representations specific to downstream scenarios, i.e., global representations suitable for tasks such as classification or local representations for tasks such as detection and localization. While they produce satisfactory results in the intended downstream scenarios, they often fail to generalize to tasks that they were not originally designed for. In this work, we propose to learn video representations that generalize to both the tasks which require global semantic information (e.g., classification) and the tasks that require local fine-grained spatio-temporal information (e.g., localization). We achieve this by optimizing two contrastive objectives that together encourage our model to learn global-local visual information given audio signals. We show that the two objectives mutually improve the generalizability of the learned global-local representations, significantly outperforming their disjointly learned counterparts. We demonstrate our approach on various tasks including action/sound classification, lip reading, deepfake detection, event and sound localization.

1 Introduction

Recent years have seen a surge of interest in contrastive self-supervised learning (CSL) [59, 38, 35, 16] to obtain representations that generalize to various downstream scenarios. In CSL, the choice of “contrasting views” plays a crucial role because the learned representations capture information shared between different views by maximizing mutual information between them [8]. This makes it critical to design contrastive objectives with the “right” contrasting views tailored for the intended downstream scenarios [77], which has been the focus of many recent works [38, 10, 45, 74, 84, 83].

The progress made so far provides important insights for understanding how to select optimal contrasting views for a given task [77]. However, the current paradigm of designing CSL approaches specific to any intended (global or local) downstream scenarios could be suboptimal, as in the real-world case the downstream scenarios are generally unknown in advance. This not only limits the generalizability of the learned representations, evaluating the approaches solely on the intended scenarios could produce misleading conclusions. Although existing approaches achieve impressive results in their intended downstream tasks, they often fail to generalize to tasks that they were not originally designed for, e.g., as we show later in our experiments, global representations do not generalize well to tasks such as lip reading [22, 21] which require local spatio-temporal information.

Motivated by this, we take an orthogonal direction to the current CSL approaches: We aim to learn representations agnostic to the types of downstream scenarios and generalize to both the scenarios that require global representations (e.g., classification) and scenarios that require local representations (e.g., localization). We focus on learning video representations using the natural

[https://github.com/yunyikristy/global_local]
audio-visual correspondence as the primary self-supervisory signal. In this scenario, the notion of
global/local representations is intertwined in space and time; we can obtain representations that
are spatially global or local, and also temporally global or local. However, most existing video
CSL approaches optimize for only global spatio-temporal representations and demonstrate them on
audio/visual video classification tasks [60, 46, 55, 61]. Part of the difficulty here is that formulating
a contrastive objective for local representations is not straightforward because of the one-to-many
relationship in audio-visual correspondences, i.e., spatially, multiple pixel regions can contribute to
the sound in the corresponding audio, and temporally, multiple temporal slices of audio can map to a
single video frame due to sampling rate differences. This hinders the development of CSL for local
video representations useful for tasks such as sound source separation and lipreading.

In this paper, we present an approach for learning global-local video representations in the CSL frame-
work. We design two cross-modal contrastive objectives that collaboratively capture information
shared between audio and visual signals. An important aspect of our approach is the factoriza-
tion of the spatio-temporal feature space into a spatially-local/temporally-global subspace and a
spatially-global/temporally-local subspace, where each of the two contrastive objectives are defined
in, respectively; see Fig. 1. The explicit space-time factorization helps each contrastive objective
focus on capturing either spatially-local or temporally-local information and thus facilitates learning
complementary features from audio-visual correspondence more effectively than in the original
spatio-temporal space. Furthermore, we define both objectives in the multiple instance learning
framework [24, 52] to handle the one-to-many relationship between audio and visual signals. This
helps the model learn representations without knowing fine-grained audio-visual correspondence.

We evaluate our approach on various downstream tasks that need local spatio-temporal information,
i.e., lip reading [22, 21, 3], deep-fake detection [25] and audio-visual event localization [26], and also
discriminative tasks that needs global information, i.e., audio/visual video classification [71, 47, 63,
41]. We show that the same pretrained model successfully generalizes to all our scenarios without
having to re-pretrain it using different objectives and/or datasets. Furthermore, we demonstrate that
the two contrastive objectives mutually benefits each other and helps improve the generalizability
of both global and local representations. To the best of our knowledge, our work is the first to
demonstrate a CSL approach that learns video representations that generalize to both global and local
video understanding scenarios.

2 Related Work

Contrastive self-supervised learning. Contrastive learning leverages multiple views of the same
data [59], e.g., multiple perspectives within the same modality such as augmentations of the same
image, different frames/clips of a video, etc. [35, 32, 37] or perspectives from different modalities
such as RGB and depth, images/videos and text [25, 53, 75]. DIM [38] and SimCLR [16] show
that leveraging local information in contrastive learning further improves performance on image
classification. DIM [38] has been extended to multi-scale [11] and to video data [37]. However,
evaluation is still focused on “discriminative” tasks, e.g., image classification and video classification,
while there is little evidence that these models will adapt well to tasks that require local information.

Several recent advances happened in the image domain, e.g., MoCo [35, 19], BYOL [51],
SwAV [13], SimSiam [18], BarlowTwins [90]. Although evaluation is performed in both global
and local downstream scenarios such as image classification, object detection, semantic/instance
segmentation and depth estimation, this line of work focuses on evaluating generalizability of the
learned global representations rather than studying the importance of global and local information
in generalization, which is the focus of this work. They also focus on image recognition tasks only.

Audio-visual video representation learning. Learning video representations from the natural audio-
visual correspondence has been studied extensively. Most existing approaches aim to capture high-
level semantic information useful for sequence-level (global) discrimination tasks such as audio/visual
video classification [29, 46, 55, 61, 23]. Along this line of work, AVSlowFast [81] utilizes different
temporal scales of the audio and visual data, which encourages the model to capture fine-grained
temporal information. However, their learning objective optimizes for only global representations
induced by different sampling rates and their evaluation is still focused on classification tasks.

Another line of work focuses on capturing fine-grained spatio-temporal local information suitable for
“local” task such as sound source separation and localization [7, 69, 93, 28, 64, 50]. Lin et
We consider $z$ and factorize the feature space into a spatially-local/temporally-global (S-local/T-global) subspace and perform temporal average pooling to obtain $z$ where $T$. To help our model capture slowly changing yet spatially-detailed information, we purposely make $\text{Spatially Local Temporally Global Subspace}$ to guide learning frame-level information via a spatially-aware attention pooling mechanism (Sec 3.3).

3 Approach

We factorize the feature space into a spatially-local/temporally-global (S-local/T-global) subspace and a spatially-global/temporally-local (S-global/T-local) subspace and define two cross-modal contrastive objectives in each of the subspaces. The S-local/T-global objective captures slowly changing patch-level information (Sec 3.1) while the S-global/T-local objective captures fast changing frame-level information (Sec 3.2). Furthermore, we utilize the learned patch-level information to guide learning frame-level information via a spatially-aware attention pooling mechanism (Sec 3.3).

3.1 Spatially-Local Temporally-Global Contrastive Objective

The purpose of this objective is to capture slowly changing patch-level information with high audio-visual correlation. We define an audio encoder $E^a_T$ that computes audio embeddings $z^a_T \in \mathbb{R}^{T_a \times F}$, where $T_a$ is the audio sequence length and $F$ is the number of frequency bands, and perform temporal average pooling to obtain $z^a \in \mathbb{R}^{1 \times F}$. Similarly, we define a visual encoder $E^v_T$ that computes visual embeddings $z^v_T \in \mathbb{R}^{T_v \times H \times W \times C}$, where $T_v$ is the visual sequence length and $H$, $W$, $C$ are the height, width and channel dimensions, and perform temporal average pooling to obtain $z^v \in \mathbb{R}^{1 \times H \times W \times C}$.

To help our model capture slowly changing yet spatially-detailed information, we purposively make input frames lack local temporal information by feeding them at a low sampling rate, e.g., at one eighths of the original sampling rate, which has been shown to be effective at capturing fine-grained spatial details [27]. For audio, we found that keeping the original sampling rate to be more effective.

We consider $z^a_T$ and $z^v_T$ that come from the same clip as positive pairs, while features coming from different clips are negatives. To capture local spatial information in $z^v_T$, we consider each cell of the $H \times W$ spatial grid as an instance and define pairs by coupling $z^a_T$ with $z^v_T[i], \forall i \in H \times W$. Note that not all $z^v_T[i]$ will have valid audio-visual correspondence to $z^a_T$; it is rather likely that only a few cells (e.g., the lip region of a talking person) will match the information in $z^a_T$. To resolve this...
We construct contrasting views by taking audio and visual features from the same local temporal neighborhood. Therefore, we use the learned S-local/T-global representation (Sec. 3.1) to obtain an audio-visual embedding pair.

To define the two subspaces we apply global pooling over either spatial or temporal dimensions. For the S-local/T-global subspace, we simply perform temporal average pooling because video inputs to our model are relatively of short length (3 seconds) and tend to show content recorded in a single scene. For the S-global/T-local subspace, however, spatial average pooling could be problematic because not all pixel regions will contribute to the sound in the corresponding audio signals.

3.2 Spatially-Global Temporally-Local Contrastive Objective

To capture information sensitive to temporal changes, e.g., the lip region of a talking head, we take audio features in a local temporal neighborhood. We compute audio embeddings $z_a^l \in \mathbb{R}^{T_v \times F}$ and visual embeddings $z_v^l \in \mathbb{R}^{T_v \times H \times W \times C}$ in the same manner as before but without applying temporal average pooling at the end. To help our model capture fine-grained temporal changes, unlike in the previous objective, we feed frames at a higher sampling rate and perform spatial pooling over visual embeddings to obtain $z_v^l \in \mathbb{R}^{T_v \times 1 \times 1 \times C}$.

We construct contrasting views by taking audio and visual features from the same local temporal block as positive pairs; audio-visual pairs from different temporal blocks of the same video become negatives. We do not use samples from different videos in order to encourage the shared information between modalities to be just how features vary over time.

Note that a single visual feature $z_v^l[t], t \in T_v$ maps to multiple audio features $z_a^l[t], t \in T_a$ because audio signals are typically captured at a higher sampling rate. Therefore, unlike in the previous objective, all the pairs within the same local temporal block can be considered as valid positives. There exists several ways to define a contrastive objective with multiple positives [42, 12, 70]; we opt for a simple approach that sums scores over all positive pairs, which leads to a form similar to Eqn. 1:

$$L^g = -\log \left( \frac{\sum_{z_a^l[i] \in P} F(z_a^g, z_v^g[i])}{\sum_{z_a^l[i] \in P} F(z_a^g, z_v^g[i]) + \sum_{z_v^l \in N} F(z_a^g, z_v^l)} \right)$$

where $F(z_a, z_v) = \exp(z_a^T z_v)$ measures the compatibility between $z_a$ and $z_v$, $P$ is a set of spatial grids in $z_v^g$, and $N$ is a set of negative visual instances taken from different clips, i.e., given a batch of $B$ video clips, we consider $H \times W \times (B - 1)^2$ negative pairs.

3.3 Spatial Pooling with Audio-Visual Attention

To define the two subspaces we apply global pooling over either spatial or temporal dimensions. For the S-local/T-global subspace, we simply perform temporal average pooling because video inputs to our model are relatively of short length (3 seconds) and tend to show content recorded in a single scene. For the S-global/T-local subspace, however, spatial average pooling could be problematic because not all pixel regions will contribute to the sound in the corresponding audio signals.

Therefore, we use the learned S-local/T-global representation (Sec. 3.1) to obtain an audio-visual attention map and use it to perform spatial attention pooling, as seen in Fig. 1. Specifically, we compute the dot product between $z_v^g$ and each of $z_v^g[i], \forall i \in H \times W$ to obtain an attention map indicating regions of high audio-visual correlation. We use this map to perform spatial pooling at each time step to obtain $z_v^l[t], \forall t \in T_v$. We demonstrate the effectiveness of spatial attention pooling in Table 8 and show qualitative examples in Fig. 2 for the scenario of sound source localization.

4 Experiments

Implementation details. We use 3D-ResNet [34] for our visual encoders ($E_v^g$ and $E_v^l$) and 1D-ResNet [36] for our audio encoders ($E_a^g$ and $E_a^l$), in both cases using Batch Normalization [39]. We share the weights of the two audio encoders and denote both simply by $E_a$; for the visual encoders we instead keep the weights separate in order to encourage them to capture complementary information from two different sampling rates; this can be seen as an instantiation of the SlowFast network with no
1 second video clips from the DFDC dataset [25] for 100 epochs with a batch size of 16. We evaluate
×
112
We apply random horizontal flipping as data augmentation. We concatenate the features produced
×
16
We evaluate our pretrained model on action recognition (UCF101 [71], HMDB51 [47], and Kinet-
ics400 [41]), on sound classification (ESC50 [63]), on lip reading (LRW [22] and LRS2 [21]), on
deepfake detection (DFDC [25]), and on audio-visual event localization (AVE [76]). We also conduct
Table 1 compares our approach with SOTA supervised and self-supervised methods. For LRS2,
approach with the same ResNet18 backbone outperforms SOTA supervised approaches on LRS and
and train the model on the cross-entropy loss. In both cases we train the whole model end-to-end.
We detect 68 facial landmarks in each frame using dlib [15] and use the outer eye and nose tip landmarks to align the detected face in each frame using an affine transform. Finally, an image of size 112 × 112 is cropped from the aligned face with the lip landmarks at the center, so that the lip region occupies one third of the image width. We apply random horizontal flipping as data augmentation. We concatenate the features produced by our pretrained $E_g^l$ and $E_l^g$ and feed them to a 2-layer MLP prediction head. For LRS2, we apply spatial average pooling before concatenation to preserve the temporal dimension and train the model using the CTC loss [30]. For LRW we apply spatio-temporal average pooling before concatenation to preserve the temporal dimension and train the model on the cross-entropy loss. In both cases we train the whole model end-to-end.

Datasets. Many video tasks involve human actions (e.g., action recognition), faces (e.g., deepfake),
and speech (e.g., lip reading). Pretraining our model on different datasets for different downstream
tasks could produce misleading conclusions as the model could simply pick up the biases in the data
useful for downstream tasks, e.g., Kinetics [14] contains various human actions useful for solving
action recognition and AVSpeech [26] mostly contains human speech useful for lip reading, but not
the other way around. Therefore, we pretrain our model only once for all downstream tasks using a
combination of Kinetics [14] and AVSpeech [26]. The standard choice of video dataset for pretraining
is Kinetics-400 [41] that contains 240K videos. We match the size by randomly selecting 120K video
samples from each datasets; we term it as K-AV-240K. For the ablation study, we pretrain our model
on a subset of 15K samples from the K-AV-240K dataset. For fair comparisons with existing work,
we also pretrain our model on the same datasets as with state-of-the-art (SOTA) approaches.

We evaluate our pretrained model on action recognition (UCF101 [71], HMDB51 [47], and Kinetics-400 [41]),
on sound classification (ESC50 [63]), on lip reading (LRW [22] and LRS2 [21]), on
deepfake detection (DFDC [25]), and on audio-visual event localization (AVE [76]). We also conduct
qualitative analysis on sound source separation on Kinetics-Sounds [6].

4.1 Downstream Scenarios

Lip Reading. Visually recognizing a speaker’s utterance is a challenging task, e.g., lip movements for
different sounding letters can be visually similar to each other (e.g., /b and /p/, /d and /t). This requires
visual representations to capture fine-grained spatio-temporal information. For a fair comparison
with SOTA, we use the standard data processing protocol of [21]. We detect 68 facial landmarks in
each frame using dlib [15] and use the outer eye and nose tip landmarks to align the detected face in
each frame using an affine transform. Finally, an image of size 112 × 112 is cropped from the aligned face with the lip landmarks at the center, so that the lip region occupies one third of the image width. We apply random horizontal flipping as data augmentation. We concatenate the features produced by our pretrained $E_g^l$ and $E_l^g$ and feed them to a 2-layer MLP prediction head. For LRS2, we apply spatial average pooling before concatenation to preserve the temporal dimension and train the model using the CTC loss [30]. For LRW we apply spatio-temporal average pooling before concatenation and train the model on the cross-entropy loss. In both cases we train the whole model end-to-end.

Table 1 compares our approach with SOTA supervised and self-supervised methods. For LRS2,
we report the word error rate (WER; the lower the better); for LRW, we evaluate on a 500-way
word classification task and report top-1 accuracy (the higher the better). The results show that our
approach with the same ResNet18 backbone outperforms SOTA supervised approaches on LRS and
LRW by large margins, i.e. 4.7% WER reduction on LRS2 and 5.1% accuracy improvements on LRW.
All the baseline self-supervised methods optimize for variants of global contrastive objectives and
generally perform poorly on all three datasets. Our approach outperforms all SOTA self-supervised
approaches with the same backbone and using the same pretraining dataset. These results demonstrate
the importance of capturing fine-grained spatio-temporal information necessary for lip reading.

Deepfake Detection. We observe that “deepfakes” tend to be characterized by fine-grained audio-
visual inconsistencies such as misalignment between lip motions and audio, unnatural facial and lip
appearance/movements or asymmetry between facial regions such as the left and right eyes. Detecting
such artifacts requires local spatio-temporal features. We take our pretrained model and finetune it on 1
second video clips from the DFDC dataset [25] for 100 epochs with a batch size of 16. We evaluate
performance using video-wise Area Under the Curve (AUC). We follow the same data preprocessing...
Table 1: Comparison with SOTA on lipreading: 
Table 2: Comparison with SOTA on deepfake detection: ⩾ ⩾ use audio-visual signals; all the other methods use visual signals only.

Table 3: Comparison with SOTA on action classification (UCF101 [71], HMDB51 [47]) and sound classification (ESC50 [63]). We highlight the best results and the second best results.

protocol as in SOTA approaches for this task, and use the same training and test sets as [20]. We perform face detection to crop the face region in each video frame. We concatenate the features produced by our pretrained $E_v$ and $E_g$ after applying spatio-temporal average pooling, feed them to a 2-layer MLP prediction head, and train the whole network end-to-end using the cross-entropy loss.

Table 2 shows the results. Two of the baselines, [20] and [54], use both visual and audio features; all the other methods use only the visual features. We can see that when using only the visual features, our approach outperforms all previous SOTA approaches (AUC=96.7). We also compare our model with SOTA self-supervised approaches. Again, the baseline self-supervised methods perform poorly on this task which require local spatio-temporal information. Our model outperforms the best self-supervised result, highlighted in blue, by a large margin (90.1 vs. 85.3).

Audio-Visual Event Localization. An "audio-visual event" is defined as an event that is both visible and audible in a video segment. Audio-video event localization is usually evaluated in two settings, i.e. fully-supervised and weakly-supervised. The former aims to predict which temporal segment of a video has an audio-visual event and what category the event belongs to; the latter assumes that only a video-level event category is available and there is no temporal event boundary information during training. This is a challenging task because an event usually appears only in a small portion of a video has an audio-visual event and what category the event belongs to; the latter assumes that only a video-level event category is available and there is no temporal event boundary information during training. This is a challenging task because an event usually appears only in a small portion of
frames within a video. Detecting which temporal segment contains an audio-visual event requires fine-grained (“local”) spatio-temporal representations, especially in weakly-supervised setting, where there is no clue about temporal event boundaries.

In Table 4 we show comparisons with the SOTA on audio-visual event localization. Again, we take the same K-AV-240K-pretrained model used in the previous experiments and finetune it on the AVE dataset [76]. We concatenate the features produced by our pretrained $E_v$ and $E_a$, and feed them to a 2-layer MLP prediction head. Especially, we first utilize the $E_a$ to perform a spatial attention pooling on $E_v$ to highlight the “important” local spatial information. In this way, the final concatenated audio-visual features contain the desired spatio-temporal local information. We evaluate our model on both fully-supervised (Fully-super.) and weakly-supervised (Weakly-super.) settings. For a fair comparison, we followed the same protocol and evaluation metric as [76]. Without bells and whistles, we achieved 82.1% localization accuracy on fully-supervised setting, and 79.8% on weakly-supervised setting, which all outperforms the SOTA, i.e. PSP [94] (77.8%) (Fully-super.) and 73.5% (weakly-super.), by large margins.

**Table 4: Comparison with SOTA on audio-visual event localization.**

| Method | Fully-sup. | Weakly-sup. |
|--------|------------|-------------|
| AVEL [76] | 68.6 | 66.7 |
| AVSDN [49] | 72.6 | 67.3 |
| CMAN [87] | 73.3 | 70.4 |
| DAM [80] | 74.5 | - |
| AVRB [66] | 74.8 | 68.9 |
| AVIN [65] | 75.2 | 69.4 |
| AVT [51] | 76.8 | 70.2 |
| CMRA [86] | 77.4 | 72.9 |
| PSP [94] | 77.8 | 73.5 |
| Ours | **82.1** | **79.8** |

**Sound Source Localization.** To further demonstrate our approach achieving good audio-visual localization, we visualize audio-visual spatial attention maps on Kinetics-Sounds [6] (see Fig.4), which contains videos deemed to have high audio-visual correspondence. Such visualization can also be considered as performing sound source localization, i.e. locate objects that making sound. To plot the figure, we use the learned audio-visual attention map, add a softmax layer and apply bilinear interpolation of the $16 \times 16$ attention map back to the original image size, i.e. $192 \times 192$. The figure shows that our learned attention maps successfully localize sounding sources in videos, especially when visual content is highly related to the corresponding audio signal. For example, the first row (video frames from “playing instruments”) shows that our model can successfully localize the sounding region. For other activities like “baby talking,” “playing basketball,” “running,” our model successfully highlights regions with humans. However, we find that the attention map incorrectly highlights regions on videos that have ambiguous audio-visual relation. We show example failure cases in the last two columns of the third row: There is no visual content that clearly relates with the audio signal, and thus the model fails to find sounding sources.

**Action and Sound Classification.** To evaluate the effectiveness of the learned global spatio-temporal representations, we evaluate our approach on action and sound classification. For action classification, we concatenate the visual features from $E_v$ and $E_a$ after spatio-temporal average pooling, feed them to a 2-layer MLP prediction head, and train the whole network end-to-end using the cross-entropy loss. For audio classification, we apply temporal average pooling to the audio features from $E_a$ and feed them to a 2-layer MLP prediction head, which is trained using the cross-entropy loss.

Table 3 shows the results. For a fair comparison to existing approaches, we report both the results pretrained on K-AV-240K and the results pretrained on Kinetics-400 [41] that contains 240K videos (K400-240K). We find that, while the K400-240K pretrained models leads to better performance on UCF101 and HMDB51 due to the similarities between datasets (all focus on human actions), the K-AV-240K pretrained models also achieve competitive results, outperforming all the baselines. Overall, on all three benchmarks, our approach achieves new SOTA results (91.1% on UCF101, 61.9% on HMDB51 and 80.1% on ESC50), demonstrating the effectiveness of learning global-local representations even for tasks that require global information. This suggests that optimizing for local representations also helps improve performance on classification tasks that require global information.

### 4.2 Ablation and Analysis

**Comparison of different contrastive objectives.** In Table 5 we compare various contrastive objectives on tasks that require fine-grained local spatio-temporal information. All the methods use...
Figure 2: Visualization of the learned audio-visual spatial attention maps on videos from Kinetics-Sounds \cite{6} show that they successfully locate sounding sources, e.g., musical instruments.

| Method             | Contrastive Objective | X-Modal | Spa. | Temp. | LRS2↓ | LRW↑ | DFDC↑ |
|--------------------|-----------------------|---------|------|-------|-------|------|-------|
| MoCo \cite{35}     | Momentum Contrast     | G       | G    | G     | 71.5  | 61.2 | 60.2  |
| CPC \cite{59}      | Predictive Coding     | G       | G    | G     | 66.7  | 65.3 | 67.9  |
| DPC \cite{33}      | Dense Predictive Coding | L     | G    | G     | 65.1  | 67.5 | 71.2  |
| VDIM \cite{37}     | Global-Local DIM      | L       | G    | G     | 53.2  | 70.7 | 85.3  |
| AVTS \cite{46}     | Audio-Visual Contrast | ✓      | G    | G     | 72.1  | 64.9 | 63.1  |
| AVSlowFast \cite{81}| A V + Rotation        | ✓      | G    | G     | 56.1  | 75.8 | 80.9  |
| Ours               | Global-Local AVG      | ✓      | G+L | G+L   | 47.8  | 83.7 | 90.1  |

Table 5: Comparison of different contrastive objectives. All the results are based on our implementation. X-Modal specifies the methods that use a cross-modal contrastive objective; others define contrastive objective on the visual modality only. We also indicate whether each contrastive objective optimizes for global and/or local representations in spatial (Spa.) and temporal (Temp.) dimensions.

the same backbone (3D-ResNet18) and the same pretraining dataset (K-AV-15K), and follow the same experimental protocol. The results show that MoCo \cite{35}, which is successful for image classification tasks, falls short on lip reading and deepfake detection. This suggests that the “vanilla” global contrastive objective may not be effective at tasks that require local information. We find that the contrastive objectives that explicitly optimize for local spatial representations – DPC \cite{33}, VDIM \cite{37}, and ours – generally perform well, suggesting the importance of local spatial contrastive objectives. We also see that AVSlowFast \cite{81} achieves competitive performance although it does not explicitly optimize local contrastive objectives. This can be explained by the fact that, unlike all the other methods, AVSlowFast \cite{81} and ours define two visual encoders to capture complementary temporal information (“slow” and “fast” changing temporal features). In addition to AVSlowFast, we explicitly optimize for global and local spatio-temporal representation via space-time factorization; the strong performance suggests the effectiveness of our global-local contrastive objectives.

**Importance of global-local joint contrastive objective.** To demonstrate the importance of jointly learning global-local representations, we ablate three components of our model: 1) do we need both the visual encoders \(E^g\) and \(E^l\)? 2) do we need both the contrastive objectives \(L^g\) and \(L^l\)? 3) do we need features from both the subspaces \(z^g\) and \(z^l\) for the downstream tasks? In Table 6 we evaluate various combinations of these. We can see that jointly optimizing both the contrastive objectives holistically improves representations in both the subspaces \(z^g\) and \(z^l\). What is particularly interesting is that optimizing the objectives not only helps learn the subspace that they are defined in, but it also helps learn the other subspace. For example, compared to optimizing only \(L^g\), optimizing it jointly with \(L^l\) helps also improve \(z^g\) (lines a and b). We also evaluate a variant of our approach that uses only the S-global/T-local encoder \(E^l\) (line d). Pretraining this with both the objectives
Table 6: The roles of global and local information on different benchmarks. LG means spatially-local/temporally-global and GL means spatially-global/temporally-local. V. Enc.: visual encoder setup, Obj.: contrastive objective used during pretraining, Feat.: features used in downstream tasks.

| Method                          | LRS2↓ | LRW↑ | DFDC↑ | UCF101↑ | HMDB51↑ |
|---------------------------------|-------|------|-------|---------|---------|
| Avg Pooling & Single Pos. Pair  | 70.9  | 65.3 | 67.9  | 82.3    | 57.1    |
| No Pooling & Multiple Pos. Pairs| 47.6  | 86.8 | 92.6  | 89.2    | 59.9    |

Table 7: Comparison of different methods to handle multiple positive audio-visual pairs in $\mathcal{L}_l$.

Audio window size. In our implementation, each video frame covers one-tenth of a second and roughly maps to three audio slices ($M = 3$). We therefore fix $M = 3$ to use all available audio slices without overlap between windows; here, we vary $M \in \{1, 3, 5, 7\}$ to see how that affects the performance. Table 9 shows that performance of this alternative approach drops significantly on tasks that require local information (LRS2, LRW, DFDC), while for classification tasks both approaches achieve comparable results. This suggests the importance of our formulation specifically on capturing local information.

| Window Size | LRS2↓ | LRW↑ | DFDC↑ | UCF101↑ | HMDB51↑ |
|-------------|-------|------|-------|---------|---------|
| 1           | 49.8  | 87.1 | 95.0  | 87.8    | 57.9    |
| 3           | 45.1  | **89.2** | **96.7** | **90.1** | **61.3** |
| 5           | **44.9** | 89.0 | 96.5  | 89.5    | **61.6** |
| 7           | 46.8  | 88.1 | 95.1  | 88.3    | 59.0    |

Table 9: Comparison of different window sizes ($M$) for audio-visual temporal matching.

AV spatial attention (Sec. 3.3). The learned audio-visual spatial attention map can highlight discriminative face regions useful for lip reading. In Table 8 we demonstrate the quality of our attention maps by replacing lip/face bounding boxes typically used in lip reading and deepfake detection with our attention map. We note that all SOTA approaches extract features from lip/face cropped regions using off-the-shelf detectors (which require substantial supervision on their own). First, as a baseline we evaluate a variant of our approach that is trained directly on full frames without using attention maps or lip/face detectors (“Ours/Full Frame”); the performance drops significantly on all three benchmarks. Next, we extract features from the entire frame (no lip/face 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 lipreading/deepfake detection. The results ("Ours/Attention") show that it achieves results comparable to our best setting ("Ours/Crop"), which extracts features from the cropped lip/face region similar to the SOTA approaches. Notably, the attention-based approach outperforms SOTA on LRW and DFDC even without relying on lip/face region detectors, demonstrating the effectiveness.

Pretraining on a large-scale dataset. We also investigate how the pretraining scale affects the results on various downstream tasks. To this end, we pretrain our model on AudioSet [29] and
Table 8: Evaluation of the learned audio-visual spatial attention maps. “V” uses only visual sequence and “V+A” uses both visual and audio sequence for finetuning on downstream tasks.

| Task       | SOTA Results | Ours/Full Frame | Ours/Attention | Ours/Crop |
|------------|--------------|-----------------|----------------|-----------|
| LRS2       | TM-seq2seq    | 49.8            | 71.9           | 51.2      | 45.1 / Lip Crop |
|            | DFTN          | 84.1            | 62.3           | 85.1      | 89.2 / Lip Crop |
| LRW†       | VDIM          | 85.3 (V)        | 68.1 (V)       | 95.9 (V)  | 96.7 (V) / Face Crop |
|            | MDS           | 91.5 (V+A)      |                |           | 97.1 (V+A) / Face Crop |

Table 10: Comparison to SOTA approaches that are pretrained on AudioSet [29]. “FT”: finetuning, “LN”: linear evaluation. “*: evaluation on Kinetics600

We report the results on both the linear (LN) evaluation and finetuning (FT) scenarios. We followed the experimental protocol used in our paper and used the same pretrained checkpoint for all downstream scenarios. For a fair comparison, we only show the results evaluated by AudioSet pretrained models for all the other comparing approaches. As we can see, among all the SOTA approaches, “Ours” achieves the best performance. When pretraining on a large-scale dataset, i.e. AudioSet with 2 million video clips, the performance can further be improved comparing with that pretrained on medium scale dataset, e.g. Kinetics and K-AV-240k.

5 Conclusion

We presented a contrastive approach for learning global and local spatio-temporal video representations from audio-visual correspondence. We showed that the space-time factorization leads to an effective solution for learning global-local representations. Unlike many prior work in the video self-supervised learning literature, we expand the downstream evaluation scenarios to include both “global” and “local” tasks and demonstrate that our approach successfully transfers to various tasks including lip reading, deepfake detection, audio-visual event localization, and action/sound classification.

Limitations. We leverage audio-visual correspondence to compute the spatial attention map. However, when applying to downstream tasks which contains videos with no audio, such a mechanism can not be used. In addition, the attention mechanism should also be considered along the temporal dimension. For example, certain frames might play more important roles in deepfake detection, e.g. frames with obvious artifacts, or in the case of lip reading, some frames or audio slices can give more clues for the model to recognize the word through visemes and/or phonemes. How to incorporate temporal or spatio-temporal attention is also an important future work.

Boarder Impact. Our paper studies self-supervised pretraining, with applications to tasks involving both audio and visual signals, e.g. deepfake detection, lip reading, audio-visual event localization and audio/video classification. As one of the core machine learning problems, self-supervised pretraining can enable machine learning to work better and more efficiently with less data and/or task-specific designs. Especially, audio and video signals are two key sensory signals in many real-world scenarios. In this sense, our work have broader applications in computer vision, audio/speech, bioinformatics, and human-machine intelligence. The datasets in our study are all publicly available. But every data-driven method brings the risk of learning biases in the data. Although our approach is promising in creating more safe and real environment, i.e. deepfake detection, we encourage the deployment of our method to be done with careful consideration of sensitive applications that have ethical implications.
References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pages 1–7. IEEE, 2018.

[2] 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, pages 1–1, 2018.

[3] 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.

[4] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelovic, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman. Self-supervised multimodal versatile networks. Advances in Neural Information Processing Systems, 33, 2020.

[5] H. Alwassel, D. Mahajan, L. Torresani, B. Ghanem, and D. Tran. Self-supervised learning by cross-modal audio-video clustering. arXiv preprint arXiv:1911.12667, 2019.

[6] R. Arandjelovic and A. Zisserman. Look, listen and learn. In ICCV, 2017.

[7] R. Arandjelovic and A. Zisserman. Objects that sound. In Proceedings of the European conference on computer vision (ECCV), pages 435–451, 2018.

[8] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. arXiv preprint arXiv:1902.09229, 2019.

[9] Y. M. Asano, M. Patrick, C. Rupprecht, and A. Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. arXiv preprint arXiv:2006.13662, 2020.

[10] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. arXiv preprint arXiv:1906.00910, 2019.

[11] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. In Advances in Neural Information Processing Systems, pages 15535–15545, 2019.

[12] Q. Cai, Y. Wang, Y. Pan, T. Yao, and T. Mei. Joint contrastive learning with infinite possibilities. arXiv preprint arXiv:2009.14776, 2020.

[13] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882, 2020.

[14] J. Carreira, E. Noland, C. Hillier, and A. Zisserman. A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987, 2019.

[15] D. Castelli and P. Pagano. Opendlib: A digital library service system. In M. Agosti and C. Thanos, editors, Research and Advanced Technology for Digital Libraries, pages 292–308, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020.

[17] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020.

[18] X. Chen and K. He. Exploring simple siamese representation learning. arXiv preprint arXiv:2011.10566, 2020.

[19] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. arXiv preprint arXiv:2104.02057, 2021.

[20] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. arXiv preprint arXiv:2005.14405, 2020.

[21] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3444–3453. IEEE, 2017.

[22] J. S. Chung and A. Zisserman. Lip reading in the wild. In Asian Conference on Computer Vision, 2016.
[23] S.-W. Chung, J. S. Chung, and H.-G. Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3965–3969. IEEE, 2019.

[24] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence, 89(1-2):31–71, 1997.

[25] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer. The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854, 2019.

[26] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv preprint arXiv:1804.03619, 2018.

[27] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6202–6211, 2019.

[28] C. Gan, D. Huang, H. Zhao, J. B. Tenenbaum, and A. Torralba. Music gesture for visual sound separation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10478–10487, 2020.

[29] J. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. pages 776–780, 03 2017.

[30] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, pages 369–376, 2006.

[31] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In CVPR, 2018.

[32] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In CVPR, 2018.

[33] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738, 2020.

[34] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[35] R. D. Hjelm and P. Bachman. Representation learning with video deep infomax. arXiv preprint arXiv:2007.13278, 2020.

[36] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670, 2018.

[37] L. Jing and Y. Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. arXiv preprint arXiv:1811.11387, 2018.

[38] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.

[39] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. arXiv preprint arXiv:2004.11362, 2020.

[40] D. Kim, D. Cho, and I. S. Kweon. Self-supervised video representation learning with space-time cubic puzzles. In AAAI, 2019.
[44] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[45] Q. Kong, W. Wei, Z. Deng, T. Yoshinaga, and T. Murakami. Cycle-contrast for self-supervised video representation learning. arXiv preprint arXiv:2010.14810, 2020.

[46] B. Korbar, D. Tran, and L. Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In Advances in Neural Information Processing Systems, 2018.

[47] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In ICCV, 2011.

[48] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656, 2018.

[49] Y.-B. Lin, Y.-J. Li, and Y.-C. F. Wang. Dual-modality seq2seq network for audio-visual event localization. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2002–2006. IEEE, 2019.

[50] Y.-B. Lin, H.-Y. Tseng, H.-Y. Lee, Y.-Y. Lin, and M.-H. Yang. Unsupervised sound localization via iterative contrastive learning. arXiv preprint arXiv:2104.00315, 2021.

[51] Y.-B. Lin and Y.-C. F. Wang. Audiovisual transformer with instance attention for audio-visual event localization. In H. Ishikawa, C.-L. Liu, T. Pajdla, and J. Shi, editors, Computer Vision – ACCV 2020, pages 274–290, Cham, 2021. Springer International Publishing.

[52] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. Advances in neural information processing systems, pages 570–576, 1998.

[53] A. Miech, J. Alayrac, L. Smaire, 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), pages 9876–9886, Los Alamitos, CA, USA, Jun 2020. IEEE Computer Society.

[54] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. Emotions don’t lie: An audio-visual deepfake detection method using affective cues. In Proceedings of the 28th ACM International Conference on Multimedia, pages 2823–2832, 2020.

[55] P. Morgado, N. Vasconcelos, and I. Misra. Audio-visual instance discrimination with cross-modal agreement. arXiv preprint arXiv:2004.12943, 2020.

[56] A. Nagrani, J. S. Chung, S. Albanie, and A. Zisserman. Disentangled speech embeddings using cross-modal self-supervision. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6829–6833, 2020.

[57] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. arXiv preprint arXiv:1906.06876, 2019.

[58] H. H. Nguyen, J. Yamagishi, and I. 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), pages 2307–2311. IEEE, 2019.

[59] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

[60] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In Proceedings of the European Conference on Computer Vision (ECCV), pages 631–648, 2018.

[61] M. Patrick, Y. M. Asano, R. Fong, J. F. Henriques, G. Zweig, and A. Vedaldi. Multi-modal self-supervision from generalized data transformations. arXiv preprint arXiv:2003.04298, 2020.

[62] K. J. Piczak. Environmental sound classification with convolutional neural networks. In International Workshop on Machine Learning for Signal Processing (MLSP), 2015.

[63] K. J. Piczak. ESC: Dataset for environmental sound classification. In Proceedings of the 23rd ACM international conference on Multimedia, 2015.

[64] R. Qian, H. D. Di Hu, M. Wu, N. Xu, and W. Lin. Multiple sound sources localization from coarse to fine. arXiv preprint arXiv:2007.06555, 2020.
[65] J. Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4372–4376, 2020.

[66] J. Ramaswamy and S. Das. See the sound, hear the pixels. In 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 2959–2968, 2020.

[67] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision, pages 1–11, 2019.

[68] H. B. Sailor, D. M. Agrawal, and H. A. Patil. Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification. In INTERSPEECH, 2017.

[69] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon. Learning to localize sound source in visual scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4358–4366, 2018.

[70] J. Song and S. Ermon. Multi-label contrastive predictive coding. arXiv preprint arXiv:2007.09852, 2020.

[71] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[72] T. Stafylakis and G. Tzimiropoulos. Combining residual networks with lstms for lipreading. arXiv preprint arXiv:1703.04105, 2017.

[73] C. Sun, F. Baradel, K. Murphy, and C. Schmid. Contrastive bidirectional transformer for temporal representation learning. arXiv preprint arXiv:1906.05743, 2019.

[74] J. Thoma, D. P. Paudel, and L. Van Gool. Soft contrastive learning for visual localization. Advances in Neural Information Processing Systems 33, 2020.

[75] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. arXiv preprint arXiv:1906.05849, 2019.

[76] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In Proceedings of the European Conference on Computer Vision (ECCV), pages 247–263, 2018.

[77] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning. arXiv preprint arXiv:2005.10243, 2020.

[78] J. Wang, J. Jiao, L. Bao, S. He, Y. Liu, and W. Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In CVPR, 2019.

[79] X. Weng and K. Kitani. Learning spatio-temporal features with two-stream deep 3d cnns for lipreading. arXiv preprint arXiv:1905.02540, 2019.

[80] Y. Wu, L. Zhu, Y. Yan, and Y. Yang. Dual attention matching for audio-visual event localization. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6291–6299, 2019.

[81] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer. Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740, 2020.

[82] J. Xiao, S. Yang, Y. Zhang, S. Shan, and X. Chen. Deformation flow based two-stream network for lip reading. arXiv preprint arXiv:2003.05709, 2020.

[83] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, Z. Li, and P. Luo. Detco: Unsupervised contrastive learning for object detection. arXiv preprint arXiv:2102.04803, 2021.

[84] Y. Xiong, M. Ren, and R. Urtasun. Loco: Local contrastive representation learning. arXiv preprint arXiv:2008.01342, 2020.

[85] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In CVPR, 2019.

[86] H. Xu, R. Zeng, Q. Wu, M. Tan, and C. Gan. Cross-modal relation-aware networks for audio-visual event localization. In ACM International Conference on Multimedia, 2020.

[87] H. Xuan, Z. Zhang, S. Chen, J. Yang, and Y. Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. Proceedings of the AAAI Conference on Artificial Intelligence, 34:279–286, 04 2020.
[88] K. Yang, B. Russell, and J. Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020.

[89] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[90] J. Zhontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

[91] X. Zhang, F. Cheng, and S. 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.

[92] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019.

[93] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.

[94] J. Zhou, L. Zheng, Y. Zhong, S. Hao, and M. Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021.

[95] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839. IEEE, 2017.

**Checklist**

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We are using public datasets and they are all cited in our paper.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
(c) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We discuss potential bias issues in Sec. 5.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]