MULTISCALE MULTIMODAL TRANSFORMER FOR MULTIMODAL ACTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

While action recognition has been an active research area for several years, most existing approaches merely leverage the video modality as opposed to humans that efficiently process video and audio cues simultaneously. This limits the usage of recent models to applications where the actions are visually well-defined. On the other hand, audio and video can be perceived in a hierarchical structure, *e.g.*, from audio signal per sampling time point to audio activities and the whole category in the audio classification. In this work, we develop a multiscale multimodal Transformer (MMT) that employs hierarchical representation learning. Particularly, MMT is composed of a novel multiscale audio Transformer (MAT) and a multiscale video Transformer (Li et al., 2022b). Furthermore, we propose a set of multimodal supervised contrastive objectives called audio-video contrastive loss (AVC) and intra-modal contrastive loss (IMC) that specifically align the two modalities for robust multimodal representation fusion. MMT surpasses previous state-of-the-art approaches by 7.3%, 1.6% and 2.1% on Kinetics-Sounds, Epic-Kitchens-100 and VGGSound in terms of the top-1 accuracy without external training data. Moreover, our MAT significantly outperforms AST (Gong et al., 2021) by 22.2%, 4.4% and 4.7% on the three public benchmark datasets and is 3× more efficient based on the number of FLOPs. Through extensive ablation studies and visualizations, we demonstrate that the proposed MMT can effectively capture semantically more separable feature representations from a combination of video and audio signals.

1 INTRODUCTION

Several visual recognition tasks have made tremendous progress in recent years due to the availability of massive annotated datasets and recent advances in Transformer architectures (Dosovitskiy et al., 2021). However, unlike humans that have an innate ability to combine data from various modalities, existing deep learning based approaches are largely dependent on visual cues as an information source. We believe that in order to achieve human-level perception and improve accuracy, an action recognition framework should be able to construe and rationalize information from *multiple modalities*. For instance, Fig. 1 depicts "woodpecker pecking tree" (the 1st row) and "footsteps on snow" (the 5th row) in VGGSound test set (Chen et al., 2020) where video only model, MViTV2 (Li et al., 2022b), incorrectly predicts them as "playing glockenspiel" (the 2nd row) and "female singing" (the 6th row). Combining the audio signal, our multiscale multimodal Transformer can successfully detect the sound emitting objects, *i.e.*, a woodpecker and shoes, in the 4th and 8th rows from the GradCam (Selvaraju et al., 2017) visualizations. Relying on visual information alone is not sufficient and can lead to misclassifications. In this work, we propose a unified architecture that is able to process multiple modalities for video classification.

Understanding videos essentially implies learning efficient spatio-temporal representations, which is a fundamental task in the computer vision community. A majority of earlier works primarily employ 3D convolutional models, such as C3D (Tran et al., 2015) or I3D (Carreira & Zisserman, 2017), that suffer from several shortcomings. Particularly, inductive biases like local connection, translation invariance, and a locally constrained receptive field substantially restrict the learning ability of convolutional models on huge datasets (Dosovitskiy et al., 2021). Hence, recent efforts have been heavily dependent on Transformer based architectures. Specifically, several approaches (Bertasius et al., 2021; Arnab et al., 2021) apply the Transformer architecture directly to videos, which are computationally inefficient (Arnab et al., 2021). Recently, Li et al. (2022b); Fan et al. (2021); Li

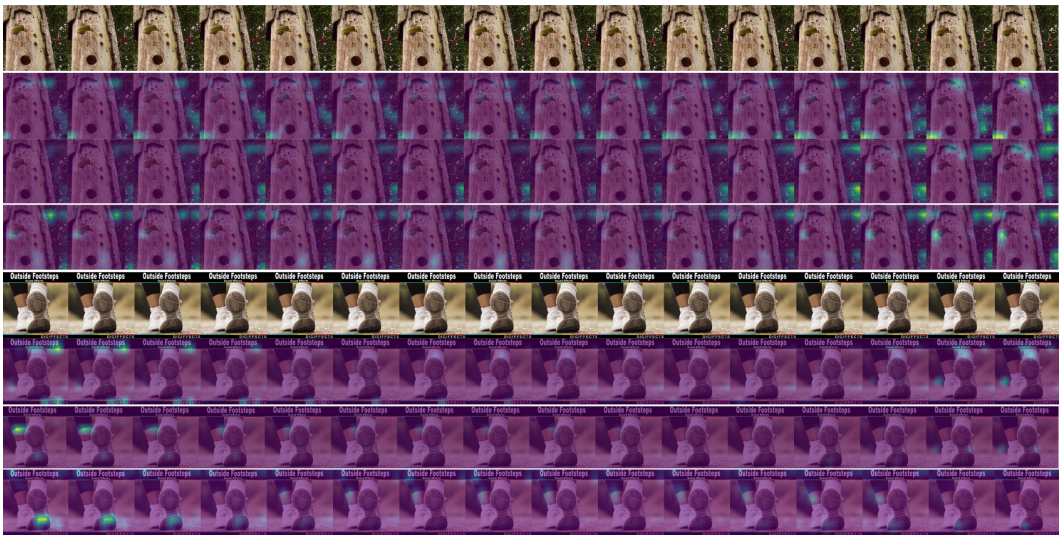

Figure 1: Two test cases, "woodpecker pecking tree" (the 1st-4th rows) and "footsteps on snow" (the 5th-8th rows), in VGGSound test set. Video only model, MViTV2 (Li et al., 2022b), incorrectly predicts them as "playing glockenspiel" and "female singing" in the 2nd and 6th rows. AVBottleneck in Sec. 3.4 incorrectly predicts the first case as "playing didgeridoo" in the 3rd row. GradCam (Selvaraju et al., 2017) visualizations in the 6th row indicate that MViTV2 mistakenly considers the text as lyrics. The 4th and 8th rows show that our MMT can align the sound with the objects, *i.e.*, a woodpecker and shoes.

et al. (2022a) propose multiscale video representation learning and pooling attention to overcome the computational cost involved and achieve the best action recognition accuracy.

On the other hand, self-supervised representation learning is prevailing, and it can fully exploit the data property and reduce the effect of inaccurate or insufficient supervised data during the training. Multimodal inputs construct multiview of each instance, and contrastive learning can be applied to multimodal signals (Zellers et al., 2022; Yang et al., 2022; Akbari et al., 2021). The self-supervised contrastive learning between multimodal signals aligns the feature embedding and enhances the multimodal fusion (Li et al., 2021).

Leveraging hierarchical representation learning from dense and simple to coarse and complex, in this work, we propose a novel multimodal Transformer to extract a joint spatio-temporal and audio representation from video and audio data sources. Specifically, the multimodal Transformer efficiently learns multiscale hierarchical representations in both audio and video encoders. To augment the learning efficiency of the multimodal Transformer, we propose a supervised multimodal alignment loss function, called audio-video contrastive (AVC) learning. The proposed loss aligns multimodal representations from the same class instead of from the same instance in previous work. Similarly, we further incorporate label supervision into intra-modality contrastive learning. Our multimodal Transformer, called multiscale multimodal Transformer (MMT), and multiscale audio Transformer (MAT) outperform previous state-of-the-art counterparts on three public datasets. Our contributions can be summarized as follows:

- We propose a novel multiscale audio Transformer (MAT), leveraging the multiscale hierarchical representation learning in audio classification. MAT progressively increases the channel capacity of the intermediate latent sequence while reducing its temporal length for audio classification.
- We construct a novel multiscale multimodal Transformer (MMT), which employs the proposed MAT and one of the current state-of-the-art video Transformers (Li et al., 2022b). To learn compact and discriminative modality representations for multimodal feature fusion, we develop audio-video contrastive loss and intra-modality contrastive loss considering label supervision.

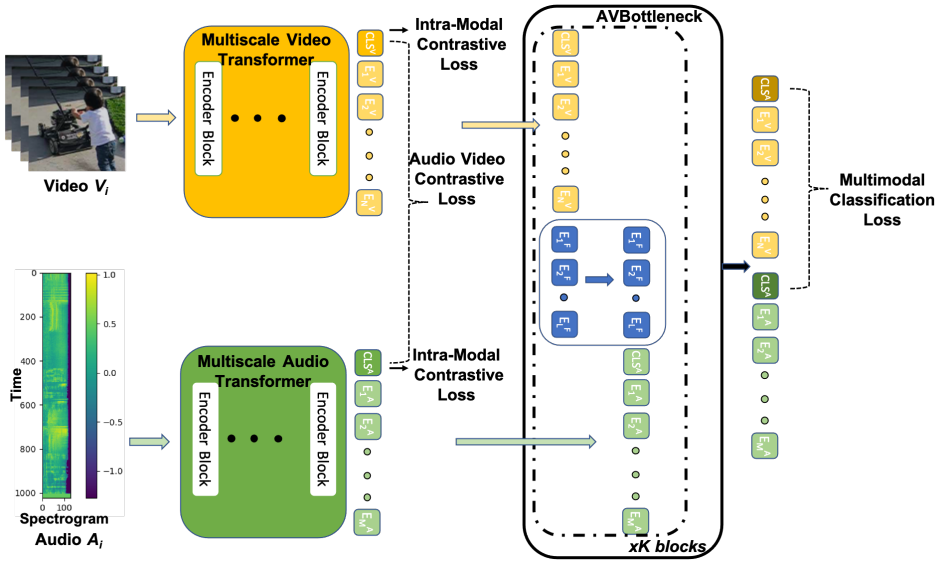

Figure 2: Framework of multiscale multimodal Transformer, MMT, where multimodal inputs are frame sequence V_i and audio spectrogram A_i from the i-th video. The multiscale audio Transformer, MAT, learns hierarchical representations, which can effectively model the temporal dependencies in audio signals since audio signal has the hierarchical structure naturally, ranging from signal at per time point to a voice activity segment and the whole audio representation. Next, we build multimodal audio-video bottleneck tokens, $\{E_1^F, \cdots, E_L^F\}$, to efficiently learn the cross-modality fusion from multiscale audio and video representations. Supervised audio-video contrastive loss and intra-modality contrastive loss encourage learning compact and discriminative representations.

- Experiments on three public datasets, Kinetics-Sounds (Arandjelovic & Zisserman, 2017), Epic-Kitchens-100 (Damen et al., 2021a) and VGGSound (Chen et al., 2020), show that MAT outperforms the previous audio Transformer (Gong et al., 2021) by 22.2%, 4.4% and 4.7%, respectively, in terms of top-1 accuracy. MMT surpasses the previous state-of-the-art counterparts by 7.3%, 1.6% and 2.1% on the three datasets without external training data.

2 Related Work

Learning effective audio-visual representations for video or audio classification can be improved by leveraging the natural alignment between audio and visual data (Owens et al., 2016; 2018; Alwassel et al., 2020; Patrick et al., 2020; Korbar et al., 2018; Chen et al., 2021b; Asano et al., 2020; Nagrani et al., 2021; Cheng et al., 2020). Moreover, audio-visual learning has several applications such as video sound localization, (Owens & Efros, 2018; Tian et al., 2018; Arandjelovic & Zisserman, 2018; Gao & Grauman, 2019; Chen et al., 2021a; Afouras et al., 2020b;a; Qian et al., 2020; Xu et al., 2020; Tzinis et al., 2021; Zhao et al., 2018; 2019), audio-visual synchronization (Ebeneze et al., 2021), person-clustering in videos (Brown et al., 2021), (visual) speech and speaker recognition (Afouras et al., 2018; Nagrani et al., 2020), and audio synthesis using visual information (Zhou et al., 2019; Goldstein & Moses, 2018; Gan et al., 2020; Koepke et al., 2020).

Video is a natural source of multimodal data. For extracting the visual features, previous approaches propose using a 3D-CNN, *e.g.*, C3D (Tran et al., 2015), R(2+1)D (Tran et al., 2018) or I3D (Carreira & Zisserman, 2017). Recently, multimodal Transformers (Nagrani et al., 2021; Akbari et al., 2021; Zellers et al., 2022) employ vision Transformer (Dosovitskiy et al., 2021) with limited number of frames, *e.g.*, eight frames, to extract visual features. For spatio-temporal representation learning, ViViT (Arnab et al., 2021) and TimeSformer (Bertasius et al., 2021) study various factorization methods along spatial- and temporal-dimensions. MViT (Fan et al., 2021; Li et al., 2022b) conducts a trade-off between resolution and the number of channels to learn a hierarchy from simple dense resolution and fine-grained features to complex coarse features. We leverage advanced multiscale hierarchy feature learning for both audio and video in our multimodal Transformer.

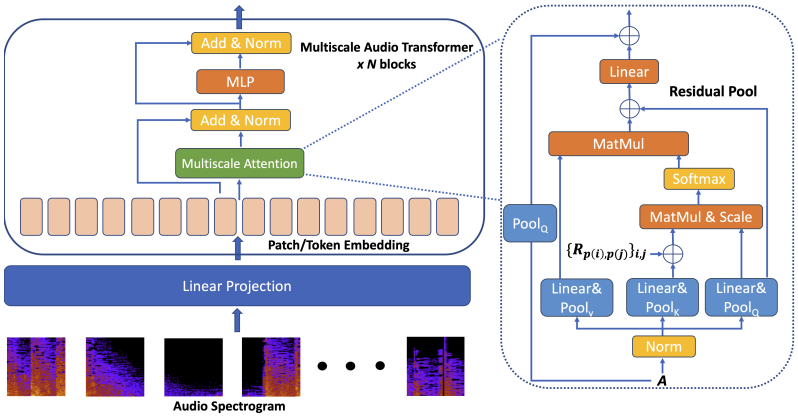

Figure 3: One block of multiscale audio Transformer (MAT). The pooling strategy in the block permits to construct representations from dense to coarse resolution and is able to effectively learn hierarchical audio representations.

Contrastive self-supervised learning can be used to align multimodal representation from different sources (Oord et al., 2018; Li et al., 2021). Yang et al. (2022) introduce intra-modality contrastive learning into multimodal fusion and obtain a better accuracy. Integrating label supervision into contrastive learning (Khosla et al., 2020) can further boost supervised learning. We enforce category discriminative cross-modality contrastive learning and intra-modality contrastive learning instead of instance discriminative contrastive learning in the multimodal Transformer.

3 Multiscale Multimodal Transformer

Multiscale multimodal Transformer, as illustrated in Fig. 2, has three main components, multiscale modality specific encoders including multiscale audio Transformer and multiscale video Transformer (Li et al., 2022b), multi-modal fusion, and multimodal learning objectives consisting of advanced audio-video contrastive loss, intra-modality contrastive loss and multimodal supervised cross-entropy loss.

3.1 Multiscale Audio Transformer

We can perceive an audio sequence in a hierarchical structure, from a signal value at each sampling time point to audio activities and an audio classification category. Therefore, hierarchical representation learning from audio spectrogram, which progressively reduces the temporal length and increases the channel dimensions, improves audio-based action recognition. Leveraging the current state-of-the-art multiscale vision Transformer (Li et al., 2022b), we construct a multiscale audio Transformer with audio spectrogram $A \in \mathbb{R}^{D\times T}$ as input, where D is the number of triangular mel-frequency bins, and T is the temporal length. The multiscale audio Transformer is illustrated in Fig. 3. One block of multiscale audio Transformer can be a stack of multihead multiscale self-attention (MMSA), layer normalization (LN) and multilayer perceptron (MLP)

$$A' = \text{MMSA}(\text{LN}(A)) + \mathcal{P}(A), \quad \text{Block}(A) = \text{MLP}(\text{LN}(A')) + A', \tag{1}$$

where $\mathcal{P}$ is a pooling operator. One attention in multihead multiscale self-attention (Li et al., 2022b) (MSAttn) can be formulated as

$$\begin{aligned} Q = \mathcal{P}_Q(AW_Q),\ K = \mathcal{P}_K(AW_K),\ V = \mathcal{P}_V(AW_V), \\ \text{MSAttn}(A) = Q + \text{Softmax}((QK^T + E^{(rel)})/\sqrt{d})V, \end{aligned} \tag{2}$$

where $E^{(rel)} = Q_i \cdot R_{p(i),p(j)} = Q_i \cdot (R^t_{t(i),t(j)} + R^f_{f(i),f(j)})$, R^t and R^f are positional embeddings along the temporal and feature axes, and d is the embedding dimension.

Compared with the previous audio spectrogram Transformer (Gong et al., 2021), multiscale audio Transformer can efficiently extract representation that effectively models hierarchical characteristics

of audio signals. In section 4, we demonstrate that the multiscale audio Transformer significantly reduces the number of parameters and FLOPs required. Using multiscale representation permits to use a larger batch size for improving the following supervised multimodal contrastive learning.

3.2 AUDIO-VIDEO CONTRASTIVE LEARNING

Multimodal inputs can naturally be considered as multiple views for the same instance in contrastive learning. Previous image-text Transformer (Li et al., 2021) shows that the image-text contrastive loss yields a better accuracy. The cross-modality contrastive learning aligns inter-modality representations, which benefits the following cross-modality fusion. The cross-modality alignment contrastive learning can be enhanced by considering label supervision to learn compact and discriminative representations.

After multiscale audio Transformer and multiscale video Transformer, we obtain audio embeddings $\{E^A_{CLS}, E^A_1, \cdots, E^A_M\}$, and video embeddings $\{E^V_{CLS}, E^V_1, \cdots, E^V_N\}$, where M is the number of audio tokens and N is the number of video tokens. The audio-video contrastive loss can be formulated

$$\mathcal{L}_{AVC} = -\mathbb{E}_{(A,V)\in D}[y_{AV} \log \frac{\exp((g_A(E^A_{CLS})^T g_V(E^V_{CLS}))/\tau)}{\sum_{(A,V)\in D} \exp((g_A(E^A_{CLS})^T g_V(E^V_{CLS}))/\tau)}], \tag{3}$$

where D is the multimodal input consisting of audio A and video V signals, y_{AV} is an indicator that the current A and V are from the same *category* or not, τ is a temperature parameter, g_A and g_V are linear embedding layers for audio representation E^A_{CLS} and video representation E^V_{CLS}, respectively. The dot product $g_A(\cdot)^T g_V(\cdot)$ measures the similarity of audio and video embedding, and the supervised audio-video contrastive learning $\mathcal{L}_{AVC}$ penalizes the distribution divergence of audio and video representations for the same *category*, which enhances the following cross-modality representation learning.

3.3 INTRA-MODALITY CONTRASTIVE LEARNING

The cross-modality fusion can also benefit from compact intra-modality representations. Yang et al. (2022) employs multiple views from data augmentation to construct intra-modality contrastive loss. We further consider label discriminative supervision into intra-modality contrastive loss

$$\begin{aligned}
\mathcal{L}^V_{IMC} &= -\mathbb{E}_{(V_1,V_2)\in D}[y_{V_1V_2} \log \frac{\exp((g_V(E^{V_1}_{CLS})^T g_V(E^{V_2}_{CLS}))/\tau)}{\sum_{(V_1,V_2)\in D} \exp((g_V(E^{V_1}_{CLS})^T g_V(E^{V_2}_{CLS}))/\tau)}], \\
\mathcal{L}^A_{IMC} &= -\mathbb{E}_{(A_1,A_2)\in D}[y_{A_1A_2} \log \frac{\exp((g_A(E^{A_1}_{CLS})^T g_A(E^{A_2}_{CLS}))/\tau)}{\sum_{(A_1,A_2)\in D} \exp((g_A(E^{A_1}_{CLS})^T g_A(E^{A_2}_{CLS}))/\tau)}],
\end{aligned} \tag{4}$$

where $y_{V_1V_2}$ and $y_{A_1A_2}$ are indicators that the current V_1 and V_2 or A_1 and A_2 are from the same *category* or not, respectively. The supervised intra-modality contrastive loss enables to learn discriminative and compact modality representations.

3.4 LEARNING FROM MULTIMODAL VIDEO

AVBottleneck Previous cross-modality Transformers either simply concatenated multimodal representations (Akbari et al., 2021), or exchanged the key and value matrices between the two modalities (Hendricks et al., 2021). However, due to the huge GPU memory consumption of the existing video Transformer, we construct an audio-video bottleneck Transformer, AVBottleneck, which handles varied lengths of modality tokens efficiently as illustrated in Fig. 2, inspired by the cross modality fusion between audio and image Transformers (Nagrani et al., 2021). Let $\{E^F_1, \cdots, E^F_L\}$ be the initial multimodal tokens, and L be the number of multimodal tokens. Without loss of generality, we omit the layer number in the denotation. One multimodal bottleneck Transformer block can be formulated as

$$\begin{aligned}
E^{VF} &= [E^V_{CLS}, E^V_1, \cdots, E^V_N, E^F_1, \cdots, E^F_L], \quad \tilde{E}^{VF} = \text{MSA}(\text{LN}(E^{VF})) + E^{VF}, \\
\hat{E}^{VF} &= \text{MLP}(\text{LN}(\tilde{E}^{VF})) + \tilde{E}^{VF}, \quad E^{AF} = [E^A_{CLS}, E^A_1, \cdots, E^A_M, \hat{E}^F_1, \cdots, \hat{E}^F_L], \\
\tilde{E}^{AF} &= \text{MSA}(\text{LN}(E^{AF})) + E^{AF}, \quad \hat{E}^{AF} = \text{MLP}(\text{LN}(\tilde{E}^{AF})) + \tilde{E}^{AF},
\end{aligned} \tag{5}$$

where multimodal tokens can be updated by averaging the multimodal tokens along all the AVBottleneck blocks. The multimodal bottleneck Transformer can be stacked into K blocks.

Models	Modal.	Kinetics-Sounds Top-1	Kinetics-Sounds Top-5	VGGSound Top-1	VGGSound Top-5
Chen et al. (2020)	A	N/A	N/A	48.8	76.5
AudioSlowFast (Kazakos et al., 2021)	A	N/A	N/A	50.1	77.9
MBT (Nagrani et al., 2021)	A	52.6	71.5	52.3	78.1
MAT (Ours)	A	**74.8** (22.2%↑)	**93.1**	**57.0** (4.7%↑)	**81.3**
AVSlowFast, R101 (Xiao et al., 2020)	A, V	85.0	N/A	N/A	N/A
MBT (Nagrani et al., 2021)	V	80.7	94.9	51.2	72.6
MBT (Nagrani et al., 2021)	A, V	85.0	96.8	64.1	85.6
MMT (Ours)	A, V	**92.3** (7.3%↑)	**99.2**	**66.2** (2.1%↑)	**85.7**

Models	Modalities	Verb	Noun	Action	FLOPs (G)
Damen et al. (2021a)	A	42.1	21.5	14.8	-
AudioSlowFast (Kazakos et al., 2021)	A	46.5	22.8	15.4	-
MBT (Nagrani et al., 2021)	A	44.3	22.4	13.0	131
MAT (Ours)	A	**50.1**	**24.2**	**17.4** (2.0%↑)	46.2
TSN (Wang et al., 2016)	V, F	60.2	46.0	33.2	-
TRN (Zhou et al., 2018)	V, F	65.9	45.4	35.3	-
TBN (Kazakos et al., 2019)	A, V, F	66.0	47.2	36.7	-
TSM (Lin et al., 2019)	V, F	67.9	49.0	38.3	-
SlowFast (Feichtenhofer et al., 2019)	V	65.6	50.0	38.5	-
MBT (Nagrani et al., 2021)	V	62.0	56.4	40.7	140
MBT (Nagrani et al., 2021)	A, V	64.8	58.0	43.4	317
ViViT-L/16×2 (Arnab et al., 2021)	V	66.4	56.8	44.0	3410
MFormer-HR (Patrick et al., 2021)	V	67.0	58.5	44.5	959
MeMViT, 16×4 (Wu et al., 2022)	V	70.6	58.5	46.2	59
MMT (Ours) (16 frames)	A, V	70.1	**61.0**	**47.8** (1.6%↑)	206

Table 1: Comparison to previous related state-of-the-art on Kinetics-Sounds (left), VGGSound (right) and Epic-Kitchens-100 (16 frames) (bottom). We report top-1 and top-5 classification accuracy on Kinetics-Sounds and VGGSound. A: Audio, V: Visual. F: Optical flow.

Computational complexity The multimodal bottleneck Transformer reduces the computing complexity from $O((M+N)^2)$ in merged concatenation based multimodal attention (Akbari et al., 2021) to $O((M+L)^2) + O((N+L)^2) \approx O(M^2) + O(N^2)$, which is the sum of complexity in one block of audio and video Transformers approximately, since $L \ll M, N$. Here, $O(M^2)$ and $O(N^2)$ are the complexities of video and audio Transformers, where M and N are the numbers of tokens in the video and audio Transformers, respectively.

Finally, we concatenate the video and audio representations $[E^V_{CLS}, E^A_{CLS}]$ and pass it through a fully connected layer for multimodal classification. The supervised multimodal loss is formulated as

$$\mathcal{L}^{AV}_{CLS} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}[y_i(c)\log p^{AV}_i(c)], \tag{6}$$

where $p^{AV}_i(c)$ is the multimodal classification probability for the i-th video and label index c. A hybrid loss consisting of multimodal video classification and supervised multimodal contrastive learning objectives forces the multimodal Transformer to learn effectively from the training data.

$$\mathcal{L} = \mathcal{L}^{AV}_{CLS} + \lambda_1\mathcal{L}_{AVC} + \lambda_2\frac{(\mathcal{L}^V_{IMC} + \mathcal{L}^A_{IMC})}{2}, \tag{7}$$

where λ_1, and λ_2 are hyperparameters to balance the loss terms in the training. The inference is consistent with the training, and we use the multimodal prediction p^{AV} directly.

Models	Top-1	Top-5
Video Only	91.6	98.8
Avg	92.0	99.1
AVBottleneck	91.2	99.1
+AL	91.4	99.0
+AVC	92.2	99.1
+AVC+IM AL	92.2	99.1
+AVC+IMC	92.3	99.2

Models	Top-1	Top-5
Video Only	56.1	77.9
Avg	62.4	84.1
AVBottleneck	63.3	84.1
+AL	63.5	84.2
+AVC	64.9	85.4
+AVC+IM AL	65.7	85.9
+AVC+IMC	66.2	85.7

Models	Video Only	Avg	AVBottleneck	+AL	+AVC	+AVC+IM AL	+AVC+IMC
Verb	67.5	68.7	69.8	69.8	70.0	69.6	70.1
Noun	59.2	59.2	59.4	59.9	60.0	60.1	61.0
Action	46.5	46.5	46.9	46.6	47.4	47.3	47.8

Table 2: Ablation study on Kinetics-Sounds (left), VGGSound (right), Epic-Kitchens-100 (bottom).

4 Experimental Results

4.1 Datasets

We experiment with three video classification datasets – Kinetics-Sounds (Arandjelovic & Zisserman, 2017), Epic-Kitchens-100 (Damen et al., 2021a; 2018; 2021b), and VGGSound (Chen et al., 2020).

Kinetics-Sounds is a commonly used subset of Kinetics (Kay et al., 2017), which consists of 10-second videos sampled at 25fps from YouTube. As Kinetics-400 is a dynamic dataset and videos may be removed from YouTube, we follow the dataset collection protocol in Xiao et al. (2020), and we collect 22,914 valid training multimodal videos and 1,585 valid test multimodal videos.

Epic-Kitchens-100 consists of 90,000 variable length egocentric clips spanning 100 hours capturing daily kitchen activities. The dataset formulates each action into a verb and a noun. We employ two classification heads, one for verb classification and the other one for noun classification. It should be noted that the dataset mainly consists of short clips with an average length of 2.6 seconds.

VGGSound is a large scale action recognition dataset, which consists of about 200K 10-second clips and 309 categories ranging from human actions and sound-emitting objects to human-object interactions. Like other YouTube datasets, *e.g.*, K400 (Kay et al., 2017), some clips are no longer available. After removing invalid clips, we collect 159,223 valid training multimodal videos and 12,790 valid test multimodal videos.

Implementation details We employ 16 frames for multiscale video Transformer and 5 ensemble views in the inference. Due to the efficiency of multiscale audio Transformer, we are able to train the model using a batch size of 64 on 8 NVIDIA A100 GPUs, each with 40 GB of memory. Following Nagrani et al. (2021), we set the numbers of AVBottleneck blocks K and tokens L as 4. τ is fixed as 0.07 and the dimensions of g_A and g_V are fixed as 256 following Li et al. (2021). For hyperparameters in multiscale audio Transformer, we follow MViTv2-B and use ImageNet-1K pretrained weights. AdamW (Loshchilov & Hutter, 2018) is used in the backpropagation and the learning rate is set as 0.0001. The number of epochs is set as 100. We set λ_1 and λ_2 as $0.25, 0.25$ for the first 20 epochs, $0.1, 0.1$ from the 21- to 40-th epochs, and $0.05, 0.05$ after the 40-th epochs. These hyperparameters are generally set to tune the loss values into the same scale. Other hyperparameters follow the recipe of MViTv2-B (Li et al., 2022b).

4.2 Results

Comparison to state-of-the-art Multiscale audio Transformer (MAT) outperforms previous audio Transformer Gong et al. (2021) by 22.2%, 4.4% and 4.7% on Kinetics-Sounds (Arandjelovic & Zisserman, 2017), Epic-Kitchens-100 (Damen et al., 2021a) and VGGSound (Chen et al., 2020) in Table 1, which demonstrates that the multiscale representation learning effectively models the

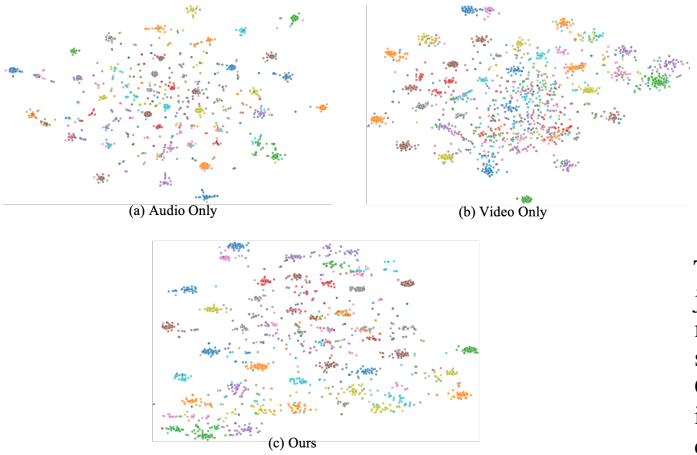

Figure 4: The t-SNE visualization (Van der Maaten & Hinton, 2008) of representations from audio only model (a), video only model (b) and our MMT (c) for random 50 categories on the test set of VGGSound.

	Video	Audio	Ours
ARI	0.394	0.370	**0.400**
HS	0.722	0.718	**0.740**

Table 3: Statistical metrics, adjusted rand index (ARI) and homogeneity score (HS), for representations of all categories on VGGSound test set. Best scores are in **bold** face. Our MMT learns a compact and discriminative representation.

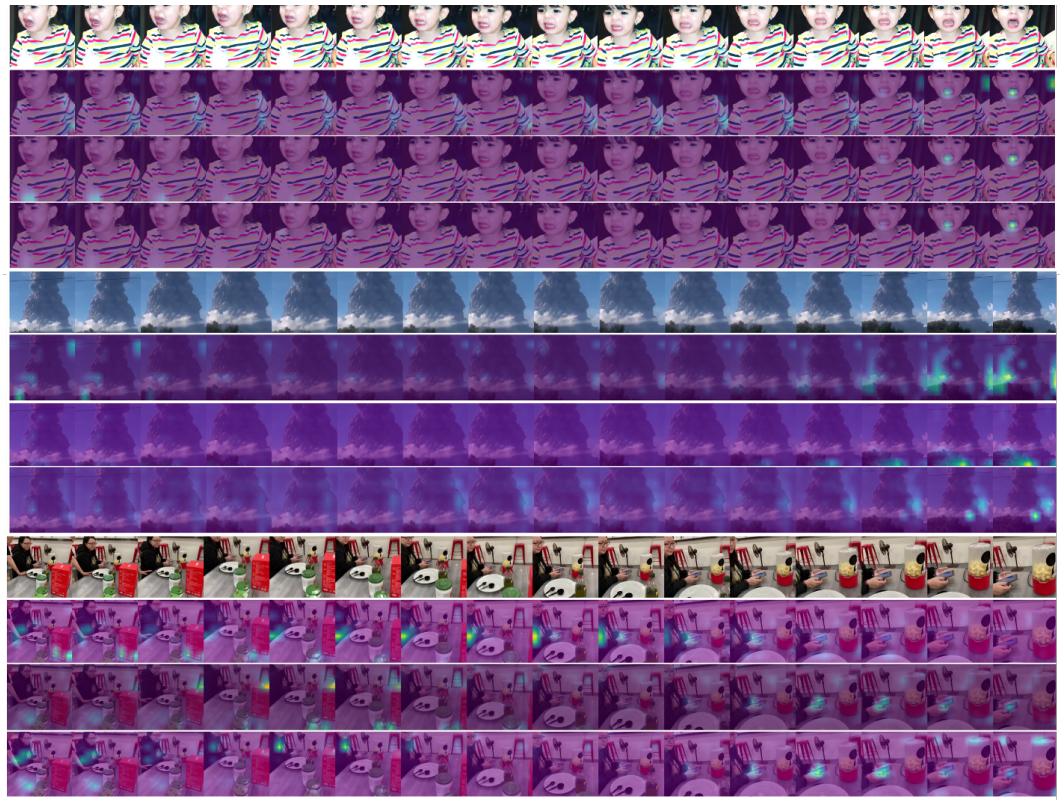

Figure 5: Visualization of three test cases in VGGSound. From top to bottom, we show 16 frames from the raw video, GradCAM (Selvaraju et al., 2017) of video only model (MViTV2), AVBottleneck, MMT (ours). With well-designed strategies to learn audio and video fusion, we demonstrate that MMT can effectively understand the clip.

hierarchical characteristics in audio signals. Multiscale multimodal Transformer (MMT) surpasses its previous state-of-the-art counterparts by 7.3%, 1.6% and 2.1% on the three public datasets based on top-1 accuracy, which shows the advantage of multiscale audio Transformer, and supervised audio-video contrastive loss and intra-modality contrastive loss. The FLOPs and #Params of our multiscale audio Transformer are 46.2G and 52M, compared with 131G FLOPs and 87M #Params in AST (Gong et al., 2021). The multiscale audio Transformer is 3$\times$ more efficient than AST, and the multiscale multimodal Transformer is 1.5$\times$ more efficient than MBT (Nagrani et al., 2021) based on the number of FLOPs.

The ablations study $w.r.t.$ video only, simple averaging audio only and video only predictions (Avg), AVBottleneck in section 3.4, with multimodal alignment loss (Li et al., 2021) (AL), $\mathcal{L}_{AVC}$ (AVC), intra-modality alignment loss (Yang et al., 2022) (IM AL) and $\mathcal{L}_{IMC}$ (IMC) are shown in Table 2 on the three datasets. From the table, we can find that 1) multimodal model outperforms one of the current state-of-the-art video Transformers (Li et al., 2022b) by a large margin, especially on VGGSound (+10.1%) and Epic-Kitchens-100 (+1.3%), 2) our multiscale multimodal Transformer with multimodal supervised contrastive learning surpasses simply fusion strategies, *i.e.*, Avg and AVBottleneck, 3) supervised multimodal contrastive losses in multimodal Transformer, *i.e.*, AVC and IMC, achieve better accuracy than their naïve contrastive learning counterparts, *i.e.*, AL and IM AL, because the supervised contrastive learning (Khosla et al., 2020) can effectively use the label supervision and learns a compact discriminative representation.

Visualizations We randomly pick three test clips with category names of "baby crying", "volcano explosion", and "popping popcorn" from VGGSound test set, and visualize 16 frames of raw video, GradCAM (Selvaraju et al., 2017) of video only model, AVBottleneck, and the fully trained multiscale multimodal Transformer (MMT) sequentially. From the first test case (the 1-4th rows), we can find the video only model focuses on the body of the baby and incorrectly predicts this clip as "people screaming". With audio signal and supervised multimodal contrastive learning, the full MMT is able to align the audio and video well, and focuses only on the mouth of the baby to obtain the correct prediction. From the second test case (the 5-8th rows), we find that AVBottleneck in the 7th row cannot capture the fog and mountain, and it incorrectly predicts the clip as "mouse clicking". From the third case (the 9-12th rows), we find that the video only model does not have any attention on the pop-corn machine and only pays attention to human and the background table, and incorrectly predicts the clip as "eating with cutlery", whereas MMT with audio signal and the advanced loss function can fully interpret the underlying action in the clip.

We also employ t-SNE (Van der Maaten & Hinton, 2008) to visualize the feature representations from the second to the last layer in multiscale audio Transformer (a), multiscale video Transformer (b), and our multiscale multimodal Transformer (c) on VGGSound dataset in Fig. 4. For clarity, we randomly choose 2,000 test samples and 50 categories in the visualization. From the figure, we can find that our MMT learns a compact and discriminative representation. In Table 3, we compare the feature representations for all the categories using two statistical metrics on VGGSound test set. The adjusted rand index (ARI) (Hubert & Arabie, 1985) computes a similarity measure between the clusters and the ground truth categories. The homogeneity score (HS) (Rosenberg & Hirschberg, 2007) checks if a cluster contains samples belonging to a single class. Both metrics can be used to evaluate the compactness and correctness of representation learning methods, and a higher value means a better model. From the table, MMT achieves the best score based on the two metrics, which validates that MMT with supervised contrastive learning can effectively learn from audio and video data sources.

5 Conclusion

In this work, we have presented an effective multiscale audio Transformer, MAT, for audio classification, as well as a multiscale multimodal Transformer, MMT, for multimodal action recognition. MMT leverages advanced multiscale Transformers, supervised audio-video contrastive loss and intra-modal contrastive objective. These supervised multimodal contrastive learning objectives enable a compact and discriminative multimodal representation learning. Experimental results demonstrate that, MAT is 3$\times$ more efficient based on the number of FLOPs, and is able to outperform Gong et al. (2021) by 22.2%, 4.4% and 4.7% based on top-1 accuracy on Kinetics-Sounds, Epic-Kitchens-100 and VGGSound. MMT surpasses its previous state-of-the-art counterparts by 7.3%, 1.6% and 2.1% on the three public datasets without external training data.

REFERENCES

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

Triantafyllos Afouras, Yuki M Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. Self-supervised object detection from audio-visual correspondence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020a.

Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020b.

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

Humam Alwassel, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems*, 2020.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017.

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

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

Yuki M. Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *Advances in Neural Information Processing Systems*, 2020.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*, volume 2, pp. 4, 2021.

Andrew Brown, Vicky Kalogeiton, and Andrew Zisserman. Face, body, voice: Video person-clustering with multiple modalities. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021a.

Yanbei Chen, Yongqin Xian, A. Sophia Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021b.

Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the ACM International Conference on Multimedia*, 2020.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 2021a. URL https://doi.org/10.1007/s11263-021-01531-2.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141, 2021b. doi: 10.1109/TPAMI.2020.2991965.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Joshua P Ebeneze, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Zongyi Liu. Detection of audio-video synchronization errors via event detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6824–6835, 2021.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

Shir Goldstein and Yael Moses. Guitar music transcription from silent video. In *Proceedings of the IEEE International Conference on Computer Vision*, 2018.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Proceedings of Interspeech*, 2021.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5492–5501, 2019.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 855–859. IEEE, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

A Sophia Koepke, Olivia Wiles, Yael Moses, and Andrew Zisserman. Sight to sound: An end-to-end approach for visual piano transcription. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

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.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. In *International Conference on Learning Representations*, 2022a.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022b.

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7083–7093, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

Arsha Nagrani, Joon Son Chung, Samuel Albanie, and Andrew Zisserman. Disentangled speech embeddings using cross-modal self-supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

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

Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Learning sight from sound: Ambient sound provides supervision for visual learning. *International Journal of Computer Vision*, 2018.

Mandela Patrick, Yuki M Asano, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. In *Advances in Neural Information Processing Systems*, 2020.

Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021.

Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420, 2007.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489–4497, 2015.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *International Conference on Learning Representations*, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pp. 20–36. Springer, 2016.

Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13587–13597, 2022.

Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of the ACM International Conference on Multimedia*, 2020.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15671–15680, 2022.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16375–16387, 2022.

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

Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 803–818, 2018.

Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.