XKD: Cross-Modal Knowledge Distillation with Domain Alignment for Video Representation Learning

Pritam Sarkar[1, 2], **Ali Etemad**[1]
[1] Queen's University, Canada
[2] Vector Institute
{pritam.sarkar, ali.etemad}@queensu.ca

Abstract

We present XKD, a novel self-supervised framework to learn meaningful representations from unlabelled videos. XKD is trained with two pseudo objectives. First, masked data reconstruction is performed to learn modality-specific representations from audio and visual streams. Next, self-supervised cross-modal knowledge distillation is performed between the two modalities through a teacher-student setup to learn complementary information. We introduce a novel domain alignment strategy to tackle domain discrepancy between audio and visual modalities enabling effective cross-modal knowledge distillation. Additionally, to develop a general-purpose network capable of handling both audio and visual streams, modality-agnostic variants of XKD are introduced, which use the same pretrained backbone for different audio and visual tasks. Our proposed cross-modal knowledge distillation improves video action classification by 8% to 14% on UCF101, HMDB51, and Kinetics400. Additionally, XKD improves multimodal action classification by 5.5% on Kinetics-Sound. XKD shows state-of-the-art performance in sound classification on ESC50, achieving top-1 accuracy of 96.5%.

1 Introduction

Self-supervised learning aims to learn meaningful representations from unlabelled data with no human supervision (Chen et al. 2020; Chen and He 2021; Misra and Maaten 2020; Grill et al. 2020; He et al. 2021). Using self-supervision, recent multimodal methods (Morgado, Vasconcelos, and Misra 2021; Alwassel et al. 2020; Morgado, Misra, and Vasconcelos 2021; Min et al. 2021; Ma et al. 2020; Sarkar and Etemad 2023) have shown great promise in learning effective representations from videos. In general, multimodal frameworks leverage the existing information in multiple data streams to learn better representations for downstream tasks toward either modality (or all). Recent audio-visual frameworks aim to perform *information sharing* between networks to further enrich the learned representations (Afouras, Chung, and Zisserman 2020; Chen et al. 2021; Ren et al. 2021; Aytar, Vondrick, and Torralba 2016; Albanie et al. 2018). However, effective knowledge sharing between audio and video is particularly challenging due to the inherent diversity, complexity, and domain-specific nature of each modality, as well as the existence of substantial domain gaps between them (Ren et al. 2021; Chen et al. 2021).

In this work, we aim to perform effective information sharing between audio and video streams to obtain more generalized representations for downstream tasks. To this end, we propose a novel self-supervised framework called XKD, which stands for Cross-modal Knowledge Distillation. Our approach consists of two sets of pseudo-tasks, (*i*) masked data modelling and (*ii*) cross-modal knowledge distillation. The former is performed to learn modality-specific (MS) information, while the latter distills and transfers knowledge across modalities to further enrich the learned representations. To allow for stable and effective information exchange between modalities, we introduce a domain alignment strategy. The proposed strategy involves 2 steps (*i*) feature refinement that identifies *'what to transfer'* based on cross-modal feature relevance and (*ii*) minimizing domain discrepancy to align the two representations. Additionally, we introduce modality-agnostic (MA) variants of our method to tackle the challenging task of learning from both modalities using a unified network (shared backbone). Modality-agnostic methods are particularly useful given their ability to accept different data streams and solve a variety of tasks using the same backbone. Moreover, they ease the challenge of designing individual models for each modality, thus attracting attention in recent studies (Girdhar et al. 2022, 2023; Akbari et al. 2021).

Inspired by the recent success of Transformers in different domains (Devlin et al. 2018; Gong, Chung, and Glass 2021a; He et al. 2021; Feichtenhofer et al. 2022), we use ViT (Dosovitskiy et al. 2020) as the backbone of our framework for both audio and visual modalities. We use the 3 different sizes of datasets to pretrain the framework, including AudioSet, Kinetics400, and Kinetics-Sound. The pretrained backbones are then evaluated on multiple datasets on a variety of downstream tasks. More specifically, UCF101, HMDB51, and Kinetics400 are used for video-related tasks; ESC50 and FSD50K are used for audio-related tasks; and Kinetics-Sound is used for multimodal evaluation.

Our contributions are summarized as follows:

- We introduce XKD, a self-supervised framework for video representation learning, which uses a novel domain alignment strategy to enable self-supervised cross-modal knowledge distillation between audio and video as two heterogeneous modalities. The proposed domain alignment strategy

minimizes the domain gap and identifies the most transferable features between the two domains for effective cross-modal knowledge distillation.
- Rigorous experiments and thorough ablations are performed to analyse the proposed method. XKD achieves state-of-the-art or competitive performance on a variety of downstream tasks including video action recognition, sound classification, and multimodal fusion.
- Our proposed modality-agnostic variants achieve very competitive performance compared to the modality-specific ones, enabling the use of a single pretrained encoder for a variety of audio and visual downstream tasks.

The code, pretrained models, and supplementary material are available on the project website[1].

2 Related Work

Video self-supervised learning. Self-supervised learning has been widely used in a variety of different areas, including the challenging task of video representation learning (Schiappa, Rawat, and Shah 2022). Several prior works have attempted to learn video representations through both uni-modal (Feichtenhofer et al. 2021; Qian et al. 2021; Jing et al. 2018) as well as multimodal (Morgado, Vasconcelos, and Misra 2021; Alayrac et al. 2020; Recasens et al. 2021; Xiao, Tighe, and Modolo 2022; Han, Xie, and Zisserman 2020) pretraining. For example, prior works have explored self-supervised frameworks for learning video representations through contrastive (Morgado, Vasconcelos, and Misra 2021; Morgado, Misra, and Vasconcelos 2021; Alayrac et al. 2020), non-contrastive (Sarkar and Etemad 2023; Recasens et al. 2021), and deep clustering (Alwassel et al. 2020; Asano et al. 2020) techniques.

Masked data modelling. Inspired by the success of BERT (Devlin et al. 2018) in natural language processing, several prior works have attempted to learn meaningful representations through the reconstruction of masked (corrupted) inputs. Such methods employ encoder-decoder setups, where the encoder compresses the inputs into a lower dimension, while the decoder is trained for reconstruction. This simple approach shows promise in different domains including image (Bao, Dong, and Wei 2021; He et al. 2021; Bachmann et al. 2022), video (Wang et al. 2021; Tong et al. 2022; Feichtenhofer et al. 2022), and audio (Niizumi et al. 2022b; Gong, Chung, and Glass 2021a; Chong et al. 2022) among others.

Cross-modal knowledge distillation. The main goal of cross-modal knowledge distillation is to transfer knowledge across different modalities (Afouras, Chung, and Zisserman 2020; Dai, Das, and Bremond 2021; Chen et al. 2021; Ren et al. 2021; Aytar, Vondrick, and Torralba 2016; Albanie et al. 2018; Piergiovanni, Angelova, and Ryoo 2020). For example, (Afouras, Chung, and Zisserman 2020; Ren et al. 2021) attempted knowledge distillation between audio teachers and video students to improve visual representations. Cross-modal knowledge distillation amongst different visual modalities has been performed in (Dai, Das, and Bremond 2021), where a pretrained optical flow teacher is used to improve RGB student's representation.

[1]https://pritamqu.github.io/XKD

Modality-agnostic networks. While modality-specific models have been the preferred approach toward developing representation learning solutions due to their strong performance, they do not have the ability to learn from multiple modalities using the same backbone, which makes them harder to develop. Recently, (Girdhar et al. 2022; Akbari et al. 2021) introduced modality-agnostic models capable of learning different modalities using the same backbones. In our paper, we attempt to develop modality-agnostic variants of our solution which can handle 2 very different modalities, i.e., audio and video to solve a variety of downstream tasks.

3 Method

3.1 Overview

Figure 1 presents an overview of our framework. Our method consists of two sets of autoencoders for audio (θ^a_{ae}) and video (θ^v_{ae}). First, we train these autoencoders to reconstruct from the masked inputs, which helps the encoders (θ^a_e and θ^v_e) to learn modality-specific information. Next, to transfer knowledge between modalities to learn complementary information, we align the two domains by identifying the most transferable features and minimizing the domain gaps. Finally, audio (θ^a_t) and video (θ^v_t) teachers are employed to provide cross-modal supervision to the opposite modalities. The details of our proposed method are mentioned below.

3.2 Data Embedding

Let be given video clip x, where the visual frames and audio spectrograms are denoted as x_v and x_a, respectively. We create 'global' and 'local' views from each modality which are then used by the teachers and students respectively. First, we apply augmentations on x_v and x_a to generate the global views as x^g_v and x^g_a. Next, we generate n local views x^l_v and x^l_a from x_v and x_a respectively, where $x^l_v = \{x^l_{v1}, \ldots, x^l_{vn}\}$ and $x^l_a = \{x^l_{a1}, \ldots, x^l_{an}\}$. Specifically, x^l_{ai} is an augmented time-frequency crop of x_a and x^l_{vi} is an augmented spatio-temporal crop of x_v. To further elaborate, x^l_v and x^l_a are differently augmented than x^g_v and x^g_a. Both local and global views of audio and visual inputs are projected into an embedding space. For example, $x_a \in \mathbb{R}^{F \times T_a}$ are reshaped into N_a smaller patches of size $f \times t_a$, where $N_a = F/f \times T_a/t_a$. Similarly, we reshape the videos $x_v \in \mathbb{R}^{T_v \times H \times W \times C}$ into N_v smaller cuboids of size $t_v \times h \times w \times C$, where $N_v = T_v/t_v \times H/h \times W/w \times C$. Finally, the spectrogram patches and visual cuboids are flattened into vectors and linearly projected onto the embedding space, which are then fed to the encoders.

3.3 Masked Data Modelling

Inspired by the recent success and scalability of pretraining with masked reconstruction in different domains (Devlin et al. 2018; Bao, Dong, and Wei 2021; Wang et al. 2021; Gong, Chung, and Glass 2021a; Baevski et al. 2022; He et al. 2021; Niizumi et al. 2022b; Tong et al. 2022), we adopt masked data modelling in our framework to learn modality-specific representations. The masked reconstruction employs an autoencoder θ_{ae}, which consists of an encoder θ_e and a decoder θ_d. Let x be the input, which can be further expressed

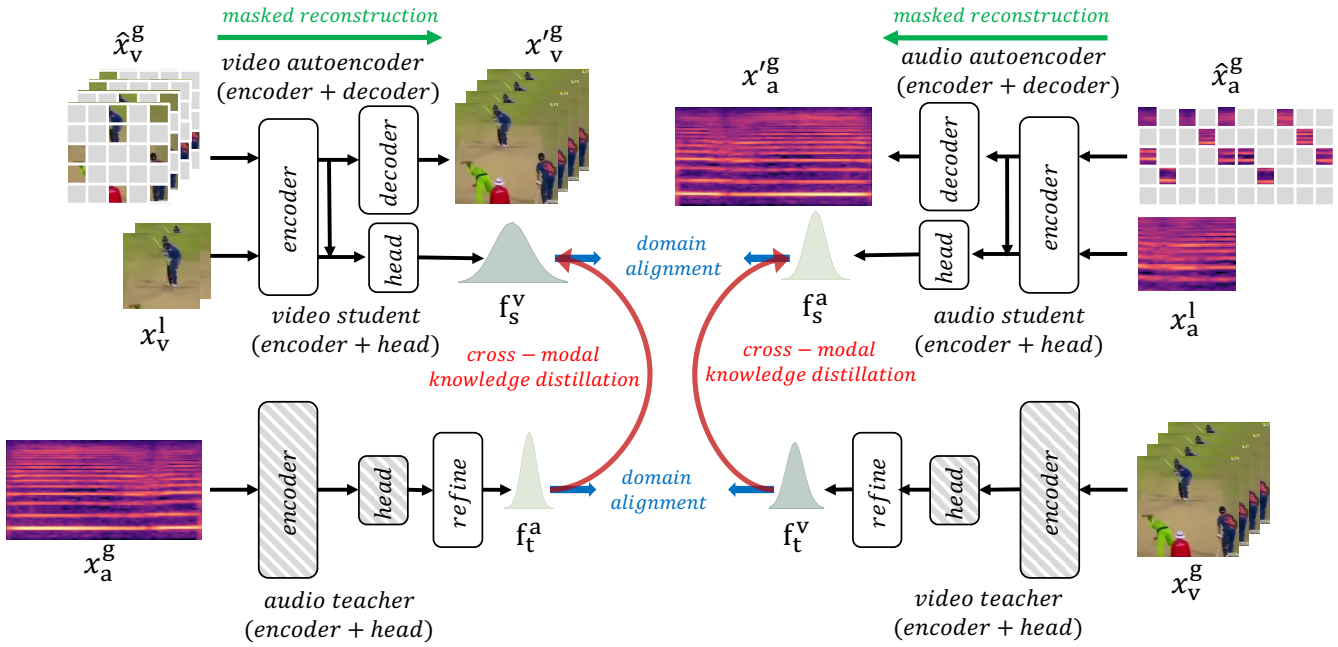

Figure 1: An overview of our proposed framework, which consists of 3 steps. Masked reconstruction: autoencoders are used to learn representations from individual modalities through reconstruction of highly masked inputs. Domain alignment: To enable cross-modal knowledge distillation, two domains are aligned through feature refinement and minimizing domain discrepancies. Cross-modal knowledge distillation: The students are used to distil knowledge from their respective cross-modal teachers.

as a sequence of tokens $\{x_i\}_{i=1}^N$, as described in Section 3.2. Here, we randomly mask some of the input tokens with $m \in \{0,1\}^N$, hence the masked tokens $x^{[m]}$ are represented as $\{x_i | m_i = 1\}_{i=1}^N$ while the corrupted inputs $\hat{x}$ are represented as $\{x_i | m_i = 0\}_{i=1}^N$. Further, we drop the masked tokens $x^{[m]}$ before feeding the input to θ_{ae} for computational efficiency (Akbari et al. 2021; He et al. 2021). We train θ_{ae} to reconstruct x, based on input $\hat{x}$. Here, θ_{ae} is trained to minimize the reconstruction loss $\mathcal{L}_{\text{recon}}$ as:

$$\mathcal{L}_{\text{recon}}(\theta_{ae}(\hat{x}), x^{[m]}) = \frac{1}{N_m} \sum_{i=1}^{N_m} (\theta_{ae}(\hat{x}_i) - x_i^{[m]})^2. \quad (1)$$

In particular, using Equation 1, we define the video and audio reconstruction losses as $\mathcal{L}^v_{\text{recon}}$ and $\mathcal{L}^a_{\text{recon}}$ for given inputs x^g_v and x^g_a, to train θ^v_{ae} and θ^a_{ae}, respectively. Here θ^v_{ae} and θ^a_{ae} denote video and audio autoencoders. To jointly train the audio-visual masked autoencoder, we define the final reconstruction loss $\mathcal{L}_{\text{ae}}$ as:

$$\mathcal{L}_{\mathbf{ae}} = \mathcal{L}_{\mathbf{recon}}(\theta^{\mathbf{v}}_{\mathbf{ae}}(\hat{\mathbf{x}}^{\mathbf{g}}_{\mathbf{v}}), \mathbf{x}^{\mathbf{g[m]}}_{\mathbf{v}}) + \mathcal{L}_{\mathbf{recon}}(\theta^{\mathbf{a}}_{\mathbf{ae}}(\hat{\mathbf{x}}^{\mathbf{g}}_{\mathbf{a}}), \mathbf{x}^{\mathbf{g[m]}}_{\mathbf{a}}). \quad (2)$$

3.4 Knowledge Distillation w/ Domain Alignment

To facilitate cross-modal knowledge sharing, we adopt a teacher-student setup (Tarvainen and Valpola 2017). The teacher (θ_t) and student (θ_s) are comprised of a backbone and a projector head (θ_h), where the teacher and student network architectures are the same but differently parameterized. Moreover, we parameterize θ_s as $\{\theta_e, \theta_h\}$, where θ_e is the same encoder used in reconstruction (explained in the previous subsection) and θ_h is a newly added projector head. We define θ^v_s and θ^a_s as video and audio students, whereas, θ^v_t and θ^a_t are denoted as video and audio teachers.

As mentioned earlier, cross-modal knowledge distillation between audio and video is particularly difficult due to the inherent diversity, complexity, and domain-specific nature of these two modalities. To tackle this, we perform domain alignment by identifying the most transferable features and minimizing the domain discrepancy. The proposed domain alignment ensures meaningful target distributions are set by the teachers in order to perform successful cross-modal knowledge distillation.

Domain Alignment. Both the audio and video carry a rich and diverse set of information about the source. Therefore, first, we identify the most transferable features by re-weighting the teachers' representations based on their cross-modal feature importance through a soft-selection process. Specifically, we obtain the cross-modal attention maps to calculate the cross-modal feature importance with respect to the corresponding modalities. In order to calculate the cross-modal attention maps, we first extract the modality-specific attention maps (A) from the last attention layers of the teacher networks as:

$$A = \text{softmax}(Q^{[\text{CLS}]} \cdot \frac{K^T}{\sqrt{d}}). \quad (3)$$

Here, Q denotes the query, K is the key, and V is the value. Specifically, $A \in \mathbb{R}^{H \times N}$ is calculated as the correlation between the query (Q) embedding of the class token (CLS) and the key (K) embeddings of all the other patches or cuboids. Note, H denotes the number of attention heads and N de-

notes the number of patches or cuboids. We obtain the visual attention $A_v \in \mathbb{R}^{H \times N_v}$ and audio attention $A_a \in \mathbb{R}^{H \times N_a}$ as per Equation 3. Next, we obtain the respective cross-modal attention maps as $A_v^{\times}$ and $A_a^{\times}$ as:

$$\begin{aligned} A_v^{\times} &= \text{MeanPool}(A_v \cdot A_a^T)/\text{scale}_\text{v}; \\ A_a^{\times} &= \text{MeanPool}(A_a \cdot A_v^T)/\text{scale}_\text{a}. \end{aligned} \tag{4}$$

Here, $A_v \cdot A_a^T \in \mathbb{R}^{H \times N_v \times N_a}$ and $A_a \cdot A_v^T \in \mathbb{R}^{H \times N_a \times N_v}$, we apply MeanPool across the last dimension. Additionally, scale_v and scale_a are scaling factors, obtained as $\frac{1}{N_v}\sum_{i=1}^{N_v} A_v$ and $\frac{1}{N_a}\sum_{i=1}^{N_a} A_a$, used to re-scale the computed cross-modal attention maps back to their original range for numerical stability. We identify the most transferable features obtained from the teachers as f_t^v and f_t^a, as $\text{refine}(\theta_t^v(x_v^g), A_v^{\times})$ and $\text{refine}(\theta_t^a(x_a^g), A_a^{\times})$ respectively. Here $A_v^{\times}$ and $A_a^{\times}$ are used to re-weight the visual and audio representations respectively. We formulate refine as:

$$\text{refine}(\theta_t(x^g), A^{\times}) = \theta_t(x^g) \cdot \frac{1}{H}\sum_{h=1}^{H} A_h^{\times} \cdot \Omega, \tag{5}$$

where $A_h^{\times}$ represents the cross-modal attention of each head and Ω is a non-negative scalar defined as the ratio of prior and posterior energy, expressed as:

$$\Omega = \frac{\| \theta_t(x^g) \|_2^2}{\| \theta_t(x^g) \cdot A^{\times} \|_2^2}. \tag{6}$$

Next, to improve the knowledge transferability, we reduce the domain gaps by minimizing the Maximum Mean Discrepancy (MMD) (Gretton et al. 2006) loss, estimated as:

$$\mathcal{L}_{\text{mmd}}(\mathbb{P}_a, \mathbb{P}_v) = \| \mathbb{E}_{x_a \sim \mathbb{P}_a}[k(\cdot, x_a)] - \mathbb{E}_{x_v \sim \mathbb{P}_v}[k(\cdot, x_v)] \|_{\mathcal{H}_k}. \tag{7}$$

Here x_a and x_v are drawn from distributions $\mathbb{P}_a$ and $\mathbb{P}_v$ respectively. Additionally, $\| \cdot \|_{\mathcal{H}_k}$ is the RKHS norm (Gretton et al. 2006) and and k represents the Gaussian kernel with bandwidth σ, written as:

$$k(x_a, x_v) = \exp\left(\frac{- \| x_a - x_v \|^2}{2\sigma^2}\right). \tag{8}$$

Using Equation 7, we define domain alignment loss $\mathcal{L}_{\text{da}}$ as:

$$\mathcal{L}_{\mathbf{da}} = \mathcal{L}_{\mathbf{mmd}}(\mathbf{f_s^v}, \mathbf{f_s^a}) + \mathcal{L}_{\mathbf{mmd}}(\mathbf{f_t^v}, \mathbf{f_t^a}). \tag{9}$$

Here, f_s^a and f_s^v refer to audio and visual representations obtained from θ_s^a and θ_s^v respectively.

Knowledge Distillation. We perform knowledge distillation between modalities to provide cross-modal supervision in order to learn complementary information. We train the student networks to match the distribution of the cross-modal teacher networks. Specifically, we train θ_s^v to match the distribution of θ_t^a, while θ_s^a is trained to match the distribution of θ_t^v. Following (Caron et al. 2021), we normalize f_t^a and f_t^v with a 'mean' calculated based on the current batch statistics, which helps the teachers' outputs be more uniformly distributed. Additionally, to prevent abrupt updates of batch means, we slowly update the 'means' using an exponential moving average (EMA). Finally, we minimize the cross-entropy loss $H(a, b)$ formulated as $-a \log b$, where a and b represent the output probability distributions of θ_t and θ_s respectively. Here, the probability P over J distributions is obtained by using a softmax function with the temperature parameter τ, where $0 < \tau < 1$ used to sharpen the distribution as:

$$P(f^{(i)}) = \frac{\exp(f^{(i)}/\tau)}{\sum_{j=1}^{J} \exp(f^{(j)}/\tau)}. \tag{10}$$

Accordingly, we define the cross-modal knowledge distillation loss $\mathcal{L}_{\text{kd}}$ as:

$$\mathcal{L}_{\mathbf{kd}} = -\mathbf{P}(\mathbf{f_t^v}) \log(\mathbf{P}(\mathbf{f_s^a})) - \mathbf{P}(\mathbf{f_t^a}) \log(\mathbf{P}(\mathbf{f_s^v})). \tag{11}$$

3.5 Final Loss

To train XKD, we define the final loss function as the combination of reconstruction, domain alignment, and cross-modal knowledge distillation losses expressed as:

$$\mathcal{L}_{\mathbf{xkd}} = \lambda_{\mathbf{ae}} \times \mathcal{L}_{\mathbf{ae}} + \lambda_{\mathbf{da}} \times \mathcal{L}_{\mathbf{da}} + \lambda_{\mathbf{kd}} \times \mathcal{L}_{\mathbf{kd}}. \tag{12}$$

Here, λ_{ae}, λ_{da}, and λ_{kd} are the loss coefficients corresponding to the three loss terms, respectively. Empirically, we set λ_{ae}, λ_{da}, and λ_{kd} as 5, 1, and 1 respectively. It should be noted that $\mathcal{L}_{\text{xkd}}$ is only used to train θ_s and θ_{ae}, not θ_t. EMA (Grill et al. 2020; Tarvainen and Valpola 2017; Chen, Xie, and He 2021) is used to slowly update θ_t^a and θ_t^v as:

$$\theta_t^a \leftarrow \lambda_a \theta_t^a + (1 - \lambda_a)\theta_s^a \text{ and } \theta_t^v \leftarrow \lambda_v \theta_t^v + (1 - \lambda_v)\theta_s^v, \tag{13}$$

where λ_a and λ_v are the EMA coefficients corresponding to θ_t^a and θ_t^v. We present the proposed algorithm in Alg. 1.

3.6 Modality-agnostic Variant

We take our proposed approach a step forward and attempt to train it in a modality-agnostic fashion with the goal of developing a 'general' network capable of handling both modalities. This is a very challenging task in the context of our work given the very diverse nature of audio and video streams. We introduce two modality-agnostic variants XKD-MATS and XKD-MAS. As the name suggests, in XKD-MATS the audio and visual teachers share their backbones, and so do the audio and visual students. In the case of XKD-MAS, the audio and visual students share their backbones, while the audio and visual teachers use modality-specific backbones. Please note that in all the setups, we use the Equation 13 to update the teachers using their respective students. Moreover, given the need to reconstruct different modalities, all of the variants use modality-specific decoders and input projection layers. The rest of the setups remain the same as the original XKD. Both variants are trained with $\mathcal{L}_{\text{xkd}}$ (see Equation 12).

4 Experiments and Results

4.1 Implementation Details

Datasets. We pretrain XKD on 3 datasets of different sizes including the small-scale **Kinetics-Sound** (Arandjelovic and Zisserman 2017), large-scale **Kinetics400** (Kay et al. 2017), and very large-scale **AudioSet** (Gemmeke et al. 2017). We evaluate our proposed method on a variety of downstream tasks including video action recognition, sound classification, and multimodal action classification. We use a total of 6

Algorithm 1: XKD Training Algorithm.

Input: x_v, x_a
Initalize: θ_d^v, θ_d^a, θ_s^v, θ_s^a, θ_t^v, θ_t^a
for t **in** iterations **do**
  Obtain x_v^g and x_v^l from x_v
  Obtain x_a^g and x_a^l from x_a
  # masked data reconstruction
  $f_{sv}^g = \theta_e^v(\hat{x}_v^g)$; $f_{sa}^g = \theta_e^a(\hat{x}_a^g)$
  # total mask reconstruction loss
  $\mathcal{L}_{ae} = \mathcal{L}_{recon}(\theta_d^v(f_{sv}^g), x_v^{g[m]}) + \mathcal{L}_{recon}(\theta_d^a(f_{sa}^g), x_a^{g[m]})$
  # cross-modal knowledge distil. w/ domain alignment
  with no-gradient:
    $f_{tv}^g = \theta_t^v(x_v^g)$; $f_{ta}^g = \theta_t^a(x_a^g)$
    Obtain $A_v^\times$, $A_a^\times$ ▷ Eqn. 4
    # feature refinement
    $f_t^v = \text{refine}(f_{tv}^g, A_v^\times)$ ▷ Eqn. 5
    $f_t^a = \text{refine}(f_{ta}^g, A_a^\times)$ ▷ Eqn. 5
  $f_{sv}^l = \theta_s^v(x_v^l)$; $f_{sa}^l = \theta_s^a(x_a^l)$
  $f_s^v = \text{Concat}\,(\theta_h^v(f_{sv}^g), f_{sv}^l)$
  $f_s^a = \text{Concat}\,(\theta_h^a(f_{sa}^g), f_{sa}^l)$
  # domain alignment loss
  $\mathcal{L}_{da} = \mathcal{L}_{mmd}(f_s^v, f_s^a) + \mathcal{L}_{mmd}(f_t^v, f_t^a)$
  # cross-modal knowledge distillation loss
  $\mathcal{L}_{kd} = -P(f_t^v)\log(P(f_s^a)) - P(f_t^a)\log(P(f_s^v))$
  # final loss
  $\mathcal{L}_{xkd} = \lambda_{ae} \times \mathcal{L}_{ae} + \lambda_{da} \times \mathcal{L}_{da} + \lambda_{kd} \times \mathcal{L}_{kd}$
  Update θ_d^v, θ_d^a, θ_s^v, θ_s^a based on $\mathcal{L}_{xkd}$
  Update θ_t^v and θ_t^a ▷ Eqn. 13
end for
Output: θ_s^v, θ_s^a, θ_t^v, θ_t^a

datasets for downstream tasks, namely **Kinetics400** (K400) (Kay et al. 2017), **Kinetics-Sound** (KS) (Arandjelovic and Zisserman 2017), **UCF101** (U101) (Soomro, Zamir, and Shah 2012), **HMDB51** (H51) (Kuehne et al. 2011), **ESC50** (E50) (Piczak 2015), and **FSD50K** (F50K) (Fonseca et al. 2022). The dataset details are provided in the supplementary material (Suppl. Mat.) Sec. A.1. Unless mentioned otherwise, Kinetics400 is used for pretraining.

Input setup. During pretraining, to save computation we downsample the video input at 8 FPS and resize the frame resolution at 112^2. Additionally, we re-sample the audio waveforms at 16 kHz. and generate mel-spectrograms using 80 mel filters. Next, we create global and local views for both audio and video. We use 4 seconds of audio-visual input for the global views. Followed by the local views are generated by taking random temporal segments of 1 second unless stated otherwise. The final input sizes to the teachers are $3 \times 32 \times 112^2$ and 80×448. Similarly, the input sizes to the students are $3 \times 8 \times 96^2$ and 80×112, for video and audio respectively. Moreover, the inputs to the encoders for masked reconstructions are the same as the input to the teacher networks, except they are heavily masked. We use a patch size of 4×16 for audio spectrograms and a cuboid size of 4×16^2 for video input. Please see additional details in the Suppl. Mat. Sec. A.2 and A.3.

Architecture. We choose ViT (Dosovitskiy et al. 2020) as the backbone for both audio and video, due to its stability in performance across different data streams (Akbari et al. 2021). In particular, we experiment with two ViT variants **ViT-B** and **ViT-L**. By default, ViT-B is used as the video backbone for pretraining with Kinetics400 and Kinetics-Sound, whereas, ViT-Large is used when pretrained with AudioSet. In all the setups, ViT-B is used as the audio backbone. The additional details are in the Suppl. Mat. Sec. A.3 and A.4.

Evaluation. Following (Sarkar and Etemad 2023; Morgado, Vasconcelos, and Misra 2021; Alwassel et al. 2020; Recasens et al. 2021), we evaluate XKD in both linear and finetuning setup. We redirect readers to see the details on evaluation protocols in the Suppl. Mat. Sec. A.5 and A.6.

4.2 Effect of Cross-modal Knowledge Distillation

As discussed, the proposed XKD is pretrained to solve masked data modelling and cross-modal knowledge distillation. Therefore, to obtain an accurate understanding of the impact of cross-modal knowledge distillation, we compare XKD ($\mathcal{L}_{\mathbf{xkd}}$) with respect to the following baselines:

- masked video reconstruction ($\mathcal{L}_{\mathbf{recon}}^{\mathbf{v}}$)
- masked audio reconstruction ($\mathcal{L}_{\mathbf{recon}}^{\mathbf{a}}$)
- audio-video masked autoencoder ($\mathcal{L}_{\mathbf{ae}}$).

We pretrain the above variants for the full training schedule of 800 epochs and report linear evaluation results on the split-1 of downstream benchmarks.

Visual representations. The results presented in Table 1 show that video representations significantly benefit from the cross-modal supervision obtained through knowledge distillation. In comparison to $\mathcal{L}_{\text{recon}}^{v}$, $\mathcal{L}_{\text{xkd}}$ improves video action recognition by 8.6%, 8.2%, and 13.9% on UCF101, HMDB51, and Kinetics-Sound, respectively. Furthermore, we challenge XKD on Kinetics400 in a linear evaluation setup which improves the top-1 accuracy by 15.7%. We interestingly notice that the joint audio-visual masked reconstruction ($\mathcal{L}_{\text{ae}}$) does not make considerable improvements, e.g., it improves top-1 accuracies by only $0.1\% - 0.2\%$ on UCf101, HMDB51, and Kinetics-Sound.

Audio representations. In Table 1, we notice that cross-modal knowledge distillation improves the performance in sound classification by 1.5% and 1% on FSD50K and ESC50 respectively. We note that the improvement is relatively less prominent compared to visual representations. Our thorough literature review in this regard reveals that a similar phenomenon has also been noticed amongst earlier works (Chen et al. 2021; Ren et al. 2021) that have attempted cross-modal knowledge distillation between audio and video in a semi-supervised setting. We conclude that while audio-to-video knowledge distillation is highly effective, video-to-audio provides a less substantial improvement. This is likely since a sharper distribution is preferred to provide supervision (Caron et al. 2021; Chen, Xie, and He 2021). As shown in Figure 2, the distribution of audio is sharper while the distribution of video is quite wider in nature. We find that such constraints can be overcome to some extent by applying aggressive sharpening of visual representations. Please find additional discussions on temperature scheduling in the Suppl. Mat. Sec. B.1. Additionally, readers are redirected to the Suppl. Mat. Sec. B.2 to B.4 for studies on EMA schedule, local views, and mask ratio.

Loss	UCF101	HHDB51	Kinetics-Sound	Kinetics400	FSD50K	ESC50
$\mathcal{L}_{\mathbf{recon}}$	76.1 ($\downarrow$8.6)	51.1 ($\downarrow$8.2)	56.8 ($\downarrow$13.9)	30.7 ($\downarrow$15.7)	44.6 ($\downarrow$1.2)	90.0 ($\downarrow$1.0)
$\mathcal{L}_{\mathbf{ae}}$	76.3 ($\downarrow$8.4)	51.2 ($\downarrow$8.1)	56.9 ($\downarrow$13.8)	32.2 ($\downarrow$14.2)	44.3 ($\downarrow$1.5)	90.0 ($\downarrow$1.0)
$\mathcal{L}_{\mathbf{xkd}}$	**84.7**	**59.3**	**70.7**	**46.4**	**45.8**	**91.0**

Table 1: Effect of cross-modal knowledge distillation in video action recognition (UCF101, HMDB51, Kinetics-Sound, and Kinetics400) and sound classification (FSD50K and ESC50). We perform linear evaluation and report the top-1 accuracy for all datasets except FSD50K, for which we report the mean average precision (mAP).

Loss	UCF101	HMDB51	Kinetics400	Remarks
$\mathcal{L}_{\mathrm{ae}}$	76.3	51.2	32.2	multimodal baseline.
$\mathcal{L}_{\mathrm{ae}}+\mathcal{L}_{\mathrm{kd}}$	76.3	51.2	32.2	without $\mathcal{L}_{\mathrm{da}}$, knowledge distillation fails.
$\mathcal{L}_{\mathrm{ae}}+\mathcal{L}_{\mathrm{da}}$	77.8	51.4	33.7	without $\mathcal{L}_{\mathrm{kd}}$, $\mathcal{L}_{\mathrm{da}}$ has marginal impact.
$\mathcal{L}_{\mathrm{ae}}+\mathcal{L}_{\mathrm{kd}}+\mathcal{L}_{\mathrm{da}}$	**84.7**	**59.3**	**46.4**	$\mathcal{L}_{\mathrm{da}}+\mathcal{L}_{\mathrm{kd}}$ improves the accuracy by $8-14\%$.

Table 2: Effect of domain alignment. $\mathcal{L}_{\mathrm{da}}$ and $\mathcal{L}_{\mathrm{kd}}$ are complementary to each other. While $\mathcal{L}_{\mathrm{da}}$ and $\mathcal{L}_{\mathrm{kd}}$ are not effective when applied separately, their combined optimization significantly improves the performance.

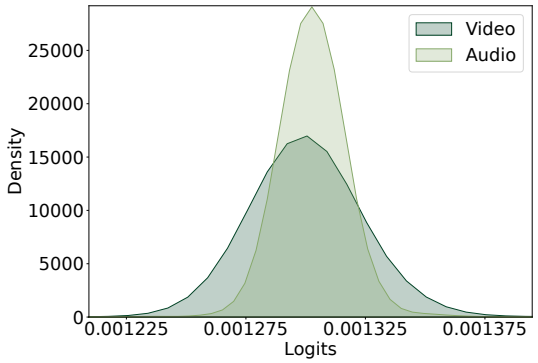

Figure 2: Visualizing the distribution of audio and visual representations, obtained from Kinetics400. It shows that audio has a sharper distribution compared to video. We find that a sharper distribution inherently provides better cross-modal supervision.

4.3 Effect of Domain Alignment

We conduct a thorough ablation study investigating the impact of domain alignment in our proposed framework, as presented in Table 2. First, without domain alignment, the model fails to perform cross-modal knowledge distillation due to domain discrepancy, and the model behaves as an audio-visual masked autoencoder. Second, quite expectedly, naively aligning the two domains without knowledge distillation has a minor impact. Last, our results exhibit that the proposed $\mathcal{L}_{da}$ and $\mathcal{L}_{kd}$ are complementary to each other, and their combined effect significantly improves the performance, e.g., by $8.1-14.2\%$ on UCF101, HMDB51, and Kinetics400. In addition to the absence of domain alignment, we identify two more factors that could cause training instability, discussed in the Suppl. Mat. Sec. B.5 and B.6. Please see alternative design choices for domain alignment in the Suppl. Mat. Sec. B.7.

	without	with
HMDB51	53.5 ($\downarrow$5.8)	**59.3**
UCF101	81.0 ($\downarrow$3.7)	**84.7**
ESC50	88.0 ($\downarrow$3.0)	**91.0**

Table 3: Effect of $\mathbf{refine}$ is presented. $\mathbf{refine}$ significantly improves downstream task performance on both audio and video tasks.

4.4 Effect of Feature Refinement ($\mathbf{refine}$)

To study the impact of refine in cross-modal knowledge distillation, we modify $\mathcal{L}_{\mathrm{da}}$ in the final loss function (Equation 12) as $\mathcal{L}_{\mathrm{mmd}}(\theta_t^a(x_a^g), \theta_t^v(x_v^g)) + \mathcal{L}_{\mathrm{mmd}}(f_s^v, f_s^a)$. The results presented in Table 3 demonstrate that refine improves downstream performance by 5.8%, 3.7%, and 3.0% on HMDB51, UCF101, and ESC50 respectively. In Figure 3, we visualize the attention from the last layer with and without refine, which further confirms the ability of refine in identifying the most transferable and key features for enhanced downstream performance. Please see alternative design choices of feature refinement Suppl. Mat. Sec. B.8.

4.5 Comparing Modality-Agnostic vs. -Specific

In Table 4, we compare the performance of modality-agnostic (MA) variants with the modality-specific (MS) one. Amongst the MA variants, XKD-MATS works slightly better on ESC50, whereas, XKD-MAS shows better performance on both UCF101 and HMDB51. The results are promising, as these variants show minor performance drops (e.g., $0.7\% - 4.1\%$) in comparison to the default XKD. Please note that such performance drops are expected as we keep the encoder size fixed for both MA and MS variants (Akbari et al. 2021). To further elaborate, while we dedicate 2 separate backbones of 87M parameters to learn audio and visual representations for the MS variant, we use just one backbone of 87M parameters for *both* audio and visual modalities

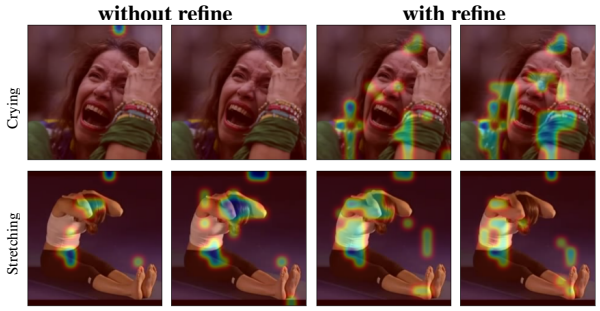

Figure 3: Visualizing the effect of refine in identifying the most transferable and key visual features. Please see more qualitative examples in the Suppl. Mat. Sec. C.1.

	MATS	MAS	MS
HMDB51	53.5(↓5.8)	55.2(↓4.1)	**59.3**
UCF101	80.3(↓4.4)	81.5(↓3.2)	**84.7**
ESC50	90.3(↓0.7)	89.5(↓1.5)	**91.0**

Table 4: Comparison between modality-agnostic (MATS, MAS) and modality-specific (MS) variants.

	Student	Teacher
HMDB51	57.8(↓1.5)	**59.3**
UCF101	84.4(↓0.3)	**84.7**
ESC50	90.3(↓0.7)	**91.0**

Table 5: Comparison between students vs. teachers.

in the MA variants. Therefore, the network parameters or model weights become saturated relatively quickly, limiting their performance. A simple solution could be to use a larger backbone for MA variants, which could be explored in future.

4.6 Comparing Teacher vs. Student

In Table 5, we present the comparison between teachers and students. Our results exhibit that due to the slow weight update schedule using EMA (see Equation 13), the teachers become slightly stronger learners compared to the students. We study the impact of different EMA schedules (presented in the Suppl. Mat. Sec. B.2) and find that EMA coefficients as 0.997 work best in our setup. As presented in Table 5, the video teacher outperforms the video student by 0.3% and 1.5% on UCF101 and HMDB51. Next, the audio teacher outperforms the audio student by 0.7% on ESC50. Please note that by default, we use the teachers to perform downstream tasks. We present a detailed comparison between the modality-specific and modality-agnostic variants using both teacher and student encoders in the Suppl. Mat. Sec. B.9.

4.7 Scalability

We study the scalability of XKD on 3 pretraining datasets of different sizes, similar to (Sarkar and Etemad 2023; Ma et al. 2020), i.e., Kinetics-Sound (22K), Kinetics400 (240K), and AudioSet (1.8M). Additionally, we experiment with 2

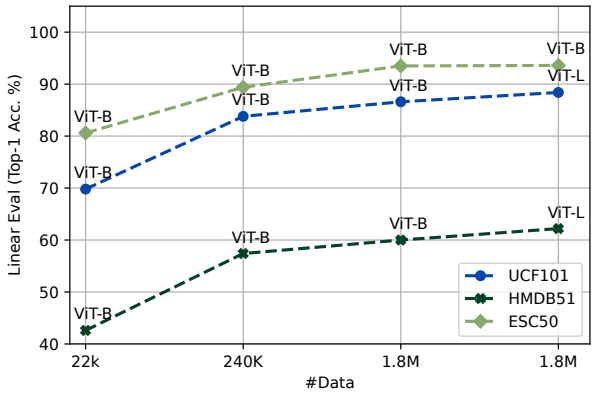

Figure 4: We study scalability both in terms of the dataset (x-axis) and network (ViT-B/ViT-L) size. Our results exhibit that XKD is scalable to both larger datasets and networks.

different sizes of ViT variants, i.e., ViT-B (87M) and ViT-L (305M). We try both ViT-B and ViT-L as the video backbone when pretrained on AudioSet. We report the linear evaluation top-1 accuracy averaged over all the splits on UCF101, HMDB51, and ESC50. Figure 4 shows that XKD continues to perform better as we increase the number of training samples and/or the size of the network. Such scalability is a much-desired property, which shows XKD can likely be scaled on even larger datasets like HowTo100M (Miech et al. 2019) or larger networks like ViT-22B (Dehghani et al. 2023). We also study the effect of longer pretraining of XKD, presented in the Suppl. Mat. Sec. B.10.

4.8 Video Action Recognition

Following the standard practice in (Alwassel et al. 2020; Sarkar and Etemad 2023; Ma et al. 2020), we compare XKD with other leading *audio-visual self-supervised* frameworks in Table 6 in both linear evaluation (**Lin.**) and finetuning (**FT.**) on UCF101, HMDB51, and Kinetics400. We note a large variability in the experiment setups amongst the prior works, however, they are included for a more inclusive and thorough comparison. We report top-1 accuracy averaged over all the splits for both UCF101 and HMDB51.

The results presented in Table 6 show that XKD outperforms or achieves competitive performance amongst the leading audio-visual self-supervised methods. For example, XKD pretrained on Kinetics400 outperforms AVID and CrissCross in linear evaluation on Kinetics400, by a very large margin. Moreover, XKD outperforms powerful state-of-the-art models like STiCA, BraVe, and CrissCross among several others when finetuned on UCF101. Additionally, XKD outperforms prior works that are pretrained with massive datasets and tri-modalities like Elo pretrained with video, audio, and optical flow from YouTube8M (YT) (Abu-El-Haija et al. 2016), compared to XKD pretrained with audio and video from AudioSet 2M. Our modality-agnostic variants also achieve encouraging results, e.g., a minor performance drop is noticed between XKD-MAS and XKD, e.g., 0.7% on UCF101 and 1.7% on Kinetics400, when finetuned.

Method	Pretrain	Modality	UCF101		HMDB51		Kinetics400	
			Lin.	FT.	Lin.	FT.	Lin.	FT.
AVSlowFast (Xiao et al. 2020)	K400	VA	77.4	87.0	44.1	54.6	-	-
SeLaVi (Asano et al. 2020)	K400	VA	-	83.1	-	47.1	-	-
XDC (Alwassel et al. 2020)	K400	VA	-	86.8	-	52.6	-	-
CMACC (Ma et al. 2020)	K400	VA	-	90.2	-	61.8	-	-
AVID (Morgado, Vasconcelos, and Misra 2021)	K400	VA	72.3*	87.5	41.4*	60.8	44.5	-
CMAC (Min et al. 2021)	K400	VA	-	90.3	-	61.1	-	-
GDT (Patrick et al. 2021a)	K400	VA	70.1*	90.9	38.5*	62.3	-	-
STiCA (Patrick et al. 2021b)	K400	VA	-	93.1	-	67.0	-	-
CrissCross (Sarkar and Etemad 2023)	K400	VA	**83.9**	91.5	50.0	64.7	44.5	-
XKD	K400	VA	83.8	**94.1**	**57.4**	**69.0**	**51.4**	**77.6**
XKD-MAS	K400	VA	81.7	93.4	55.1	65.9	50.1	75.9
XKD-MATS	K400	VA	80.1	93.1	53.1	65.7	48.8	75.7
XDC (Alwassel et al. 2020)	AS	VA	85.3	93.0	56.0	63.7	-	-
MMV (Alayrac et al. 2020)	AS	VA	83.9	91.5	60.0	70.1	-	-
CM-ACC (Ma et al. 2020)	AS	VA	-	93.5	-	67.2	-	-
BraVe** (Recasens et al. 2021)	AS	VA	**90.0**	93.6	**63.6**	70.8	-	-
AVID (Morgado, Vasconcelos, and Misra 2021)	AS	VA	-	91.5	-	64.7	48.9	-
CrissCross (Sarkar and Etemad 2023)	AS	VA	87.7	92.4	56.2	67.4	50.1	-
XKD	AS	VA	88.4	**95.8**	62.2	**75.7**	**56.5**	**80.1**
VideoMAE (Tong et al. 2022)	U101/H51	V	-	91.3	-	62.6	-	-
BEVT (Wang et al. 2021)	K400	V	-	-	-	-	-	76.2
VideoMAE (Tong et al. 2022)	K400	V	-	-	-	-	-	79.0
VideoMAE** (Tong et al. 2022)	K400	V	80.0*	96.1	54.3*	73.3	-	80.0
CPD (Li and Wang 2020)	K400	VT	-	90.5	-	63.6	-	-
CoCLR (Han, Xie, and Zisserman 2020)	K400	VF	74.5	-	46.1	-	-	-
BraVe** (Recasens et al. 2021)	AS	VFA	93.2	96.9	69.9	79.4	-	-
MIL-NCE (Miech et al. 2020)	HT	VT	-	91.3	-	61.0	-	-
VATT (Akbari et al. 2021)	AS+HT	VAT	89.6	-	65.2	-	-	81.1
VATT-MA (Akbari et al. 2021)	AS+HT	VAT	84.4	-	63.1	-	-	79.9
ELo (Piergiovanni, Angelova, and Ryoo 2020)	YT	VFA	-	93.8	-	67.4	-	-

Table 6: Comparison on video action recognition. XKD outperforms or achieves competitive performance compared to state-of-the-art methods. V: Video, A: Audio, T: Text, F: Flow. *computed by us using official checkpoints. **industry-level computation, e.g., VideoMAE uses 64 vs. ours 8 GPUs, BraVe pretrains with very high temporal resolutions (128 frames) compared to others (8-32). More comparisons with VideoMAE are in the Suppl. Mat. Sec. B.11.

4.9 Sound Classification

In Table 7, we compare the performance of our proposed method using linear evaluation (**Lin.**) and finetuning (**FT.**) on sound classification using 2 popular audio benchmarks ESC50 and FSD50K. Following (Fonseca et al. 2022; Piczak 2015), we report top-1 accuracy averaged over all the splits on ESC50 and mean average precision on FSD50K. XKD outperforms prior works like XDC, AVID, and CrissCross on ESC50 in both linear evaluation and finetuning. Moreover, XKD and its MA variants outperform VATT and VATT-MA on ESC50 by 4.7% and 7.5%, even though VATT is pretrained with 136M videos, compared to XKD which is only trained on 240K samples. Additionally, when evaluated on FSD50K, XKD outperforms BYOL-A, AudioTransformer, and PSLA among others. Lastly, XKD shows state-of-the-art performance on ESC50, achieving top-1 finetuned accuracy of 96.5% when pretrained with AudioSet.

4.10 Multimodal Fusion

Following (Sarkar and Etemad 2023), we evaluate XKD in multimodal action classification using Kinetics-Sound. We extract fixed audio and visual embeddings from the pretrained encoders and concatenate them together (i.e., late fusion), followed by a linear SVM classifier is trained and top-1 accuracy is reported. The results presented in Table 8 show that XKD outperforms CrissCross by 14.5% and baseline audio-visual masked autoencoder ($\mathcal{L}_{\mathbf{ae}}$) by 5.5%.

4.11 In-painting

We present reconstruction examples in Figure 5, which shows that XKD retains its reconstruction ability even when a very high masking ratio is applied, for both audio and video modalities. This makes XKD also suitable for in-painting tasks. More examples are in the Suppl. Mat. Sec. C.2.

Method	Pretrain	Modality	ESC50		FSD50K	
			Lin.	FT.	Lin.	FT.
XDC (Alwassel et al. 2020)	K400	VA	78.0	-	-	-
AVID (Morgado, Vasconcelos, and Misra 2021)	K400	VA	79.1	-	-	-
STiCA (Patrick et al. 2021b)	K400	VA	81.1	-	-	-
CrissCross (Sarkar and Etemad 2023)	K400	VA	86.8	-	-	-
XKD	K400	VA	**89.4**	**93.6**	**45.8**	**54.1**
XKD-MAS	K400	VA	87.3	92.7	43.0	52.3
XKD-MATS	K400	VA	88.7	92.9	43.4	53.8
AVTS (Korbar, Tran, and Torresani 2018)	AS	VA	80.6	-	-	-
AVID (Morgado, Vasconcelos, and Misra 2021)	AS	VA	89.2	-	-	-
GDT (Patrick et al. 2021a)	AS	VA	88.5	-	-	-
CrissCross (Sarkar and Etemad 2023)	AS	VA	90.5	-	-	-
XKD	AS	VA	**93.6**	**96.5**	**51.5**	**58.5**
AudioMAE (Huang et al. 2022)	AS	A	-	93.6	-	-
MaskSpec (Chong et al. 2022)	AS	A	-	89.6	-	-
MAE-AST (Baade, Peng, and Harwath 2022)	AS	A	-	90.0	-	-
BYOL-A (Niizumi et al. 2022a)	AS	A	-	-	44.8	
Aud. T-former (Verma and Berger 2021)	AS	A	-	-	-	53.7
SS-AST (Gong, Chung, and Glass 2021a)	AS	A	-	88.8	-	-
PSLA (Gong, Chung, and Glass 2021b)	AS	A	-	-	-	55.8
VATT (Akbari et al. 2021)	AS+HT	VAT	84.7	-	-	-
VATT-MA (Akbari et al. 2021)	AS+HT	VAT	81.2	-	-	-
PaSST(SL) (Koutini et al. 2021)	AS+IN	AI	-	95.5	-	58.4

Table 7: Comparison on sound classification. XKD outperforms the prior state-of-the-art methods. V: Video, A: Audio, I: Image, IN: ImageNet (Russakovsky et al. 2015).

Method	Audio + Video
CrissCross	66.7
$\mathcal{L}_{ae}$ (AV-MAE)	75.7
XKD	**81.2**
XKD-MAS	78.8
XKD-MATS	78.3

Table 8: Multimodal action classification by late fusion on Kinetics-Sound.

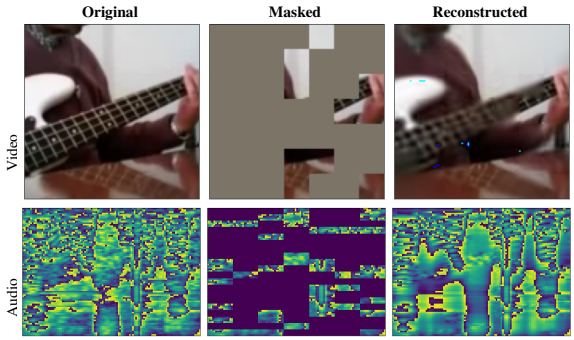

Figure 5: Reconstruction examples from highly masked inputs, video and audio mask ratios are 80% and 70%.

5 Summary

In this work, we propose XKD, a novel self-supervised framework to improve video representation learning using cross-modal knowledge distillation. To effectively transfer knowledge between audio and video modalities, XKD aligns the two domains by identifying the most transferable features and minimizing the domain gaps. Our study shows that cross-modal knowledge distillation significantly improves video representations on a variety of benchmarks. Additionally, to develop a general network with the ability to process different modalities, we introduce modality-agnostic variants of XKD which show promising results in handling both audio and video using the same backbone. We believe that our approach can be further expanded to perform cross-modal knowledge distillation between other modalities as well (e.g., vision and language), which could be investigated in the future.

Acknowledgments

We are grateful to the Bank of Montreal and Mitacs for funding this research. We are thankful to SciNet HPC Consortium for helping with the computation resources.

References

Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; and Vijayanarasimhan, S. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Afouras, T.; Chung, J. S.; and Zisserman, A. 2020. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP*, 2143–2147. IEEE.

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.-H.; Chang, S.-F.; Cui,

Y.; and Gong, B. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *NeurIPS*, 34.

Alayrac, J.-B.; Recasens, A.; Schneider, R.; Arandjelovic, R.; Ramapuram, J.; De Fauw, J.; Smaira, L.; Dieleman, S.; and Zisserman, A. 2020. Self-Supervised MultiModal Versatile Networks. *NeurIPS*, 2(6): 7.

Albanie, S.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2018. Emotion recognition in speech using cross-modal transfer in the wild. In *ACM Multimedia*, 292–301.

Alwassel, H.; Mahajan, D.; Korbar, B.; Torresani, L.; Ghanem, B.; and Tran, D. 2020. Self-Supervised Learning by Cross-Modal Audio-Video Clustering. *NeurIPS*, 33.

Arandjelovic, R.; and Zisserman, A. 2017. Look, listen and learn. In *ICCV*, 609–617.

Asano, Y. M.; Patrick, M.; Rupprecht, C.; and Vedaldi, A. 2020. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*.

Aytar, Y.; Vondrick, C.; and Torralba, A. 2016. Soundnet: Learning sound representations from unlabeled video. *NeurIPS*, 29.

Baade, A.; Peng, P.; and Harwath, D. 2022. Mae-ast: Masked autoencoding audio spectrogram transformer. *arXiv preprint arXiv:2203.16691*.

Bachmann, R.; Mizrahi, D.; Atanov, A.; and Zamir, A. 2022. MultiMAE: Multi-modal Multi-task Masked Autoencoders. *arXiv preprint arXiv:2204.01678*.

Baevski, A.; Hsu, W.-N.; Xu, Q.; Babu, A.; Gu, J.; and Auli, M. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*.

Bao, H.; Dong, L.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*, 9650–9660.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *CVPR*, 15750–15758.

Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9640–9649.

Chen, Y.; Xian, Y.; Koepke, A.; Shan, Y.; and Akata, Z. 2021. Distilling audio-visual knowledge by compositional contrastive learning. In *CVPR*, 7016–7025.

Chong, D.; Wang, H.; Zhou, P.; and Zeng, Q. 2022. Masked Spectrogram Prediction For Self-Supervised Audio Pre-Training. *arXiv preprint arXiv:2204.12768*.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 702–703.

Dai, R.; Das, S.; and Bremond, F. 2021. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *ICCV*, 13053–13064.

Dehghani, M.; Djolonga, J.; Mustafa, B.; Padlewski, P.; Heek, J.; Gilmer, J.; Steiner, A. P.; Caron, M.; Geirhos, R.; Alabdulmohsin, I.; et al. 2023. Scaling vision transformers to 22 billion parameters. In *ICML*, 7480–7512. PMLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Feichtenhofer, C.; Fan, H.; Li, Y.; and He, K. 2022. Masked Autoencoders As Spatiotemporal Learners. *arXiv preprint arXiv:2205.09113*.

Feichtenhofer, C.; Fan, H.; Xiong, B.; Girshick, R.; and He, K. 2021. A large-scale study on unsupervised spatiotemporal representation learning. In *CVPR*, 3299–3309.

Fonseca, E.; Favory, X.; Pons, J.; Font, F.; and Serra, X. 2022. FSD50K: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 829–852.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 776–780.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *CVPR*, 15180–15190.

Girdhar, R.; Singh, M.; Ravi, N.; van der Maaten, L.; Joulin, A.; and Misra, I. 2022. Omnivore: A single model for many visual modalities. In *CVPR*, 16102–16112.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021a. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021b. Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3292–3306.

Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. 2006. A kernel method for the two-sample-problem. *NeurIPS*, 19.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Pires, B.; Guo, Z.; Azar, M.; et al. 2020. Bootstrap Your Own Latent: A new approach to self-supervised learning. In *NeurIPS*.

Han, T.; Xie, W.; and Zisserman, A. 2020. Self-supervised Co-training for Video Representation Learning. In *NeurIPS*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*.

Huang, P.-Y.; Xu, H.; Li, J.; Baevski, A.; Auli, M.; Galuba, W.; Metze, F.; and Feichtenhofer, C. 2022. Masked autoencoders that listen. *NeurIPS*, 35: 28708–28720.

Jing, L.; Yang, X.; Liu, J.; and Tian, Y. 2018. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*.

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

Korbar, B.; Tran, D.; and Torresani, L. 2018. Cooperative learning of audio and video models from self-supervised synchronization. In *NeruIPS*, 7774–7785.

Koutini, K.; Schlüter, J.; Eghbal-zadeh, H.; and Widmer, G. 2021. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *ICCV*, 2556–2563.

Li, T.; and Wang, L. 2020. Learning spatiotemporal features via video and text pair discrimination. *arXiv preprint arXiv:2001.05691*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, S.; Zeng, Z.; McDuff, D.; and Song, Y. 2020. Active Contrastive Learning of Audio-Visual Video Representations. In *ICLR*.

Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; et al. 2018. Mixed Precision Training. In *ICLR*.

Miech, A.; Alayrac, J.-B.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 9879–9889.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Min, S.; Dai, Q.; Xie, H.; Gan, C.; Zhang, Y.; and Wang, J. 2021. Cross-Modal Attention Consistency for Video-Audio Unsupervised Learning. *arXiv preprint arXiv:2106.06939*.

Misra, I.; and Maaten, L. v. d. 2020. Self-supervised learning of pretext-invariant representations. In *CVPR*, 6707–6717.

Morgado, P.; Misra, I.; and Vasconcelos, N. 2021. Robust Audio-Visual Instance Discrimination. In *CVPR*, 12934–12945.

Morgado, P.; Vasconcelos, N.; and Misra, I. 2021. Audio-visual instance discrimination with cross-modal agreement. In *CVPR*, 12475–12486.

Niizumi, D.; Takeuchi, D.; Ohishi, Y.; Harada, N.; and Kashino, K. 2022a. BYOL for Audio: Exploring Pre-trained General-purpose Audio Representations. *arXiv preprint arXiv:2204.07402*.

Niizumi, D.; Takeuchi, D.; Ohishi, Y.; Harada, N.; and Kashino, K. 2022b. Masked Spectrogram Modeling using Masked Autoencoders for Learning General-purpose Audio Representation. *arXiv preprint arXiv:2204.12260*.

Patrick, M.; Asano, Y. M.; Kuznetsova, P.; Fong, R.; Henriques, J. F.; Zweig, G.; and Vedaldi, A. 2021a. Multi-modal Self-Supervision from Generalized Data Transformations. *ICCV*.

Patrick, M.; Huang, P.-Y.; Misra, I.; Metze, F.; Vedaldi, A.; Asano, Y. M.; and Henriques, J. F. 2021b. Space-Time Crop & Attend: Improving Cross-modal Video Representation Learning. In *ICCV*, 10560–10572.

Piczak, K. J. 2015. ESC: Dataset for Environmental Sound Classification. In *ACM Multimedia*, 1015–1018. .

Piergiovanni, A.; Angelova, A.; and Ryoo, M. S. 2020. Evolving losses for unsupervised video representation learning. In *CVPR*, 133–142.

Qian, R.; Meng, T.; Gong, B.; Yang, M.-H.; Wang, H.; Belongie, S.; and Cui, Y. 2021. Spatiotemporal contrastive video representation learning. In *CVPR*, 6964–6974.

Recasens, A.; Luc, P.; Alayrac, J.-B.; Wang, L.; Strub, F.; Tallec, C.; Malinowski, M.; Pătrăucean, V.; Altché, F.; Valko, M.; et al. 2021. Broaden your views for self-supervised video learning. In *ICCV*, 1255–1265.

Reddi, S. J.; Kale, S.; and Kumar, S. 2019. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.

Ren, S.; Du, Y.; Lv, J.; Han, G.; and He, S. 2021. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *CVPR*, 13325–13333.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *IJCV*, 115: 211–252.

Sarkar, P.; and Etemad, A. 2023. Self-supervised audio-visual representation learning with relaxed cross-modal synchronicity. In *AAAI*, volume 37, 9723–9732.

Schiappa, M. C.; Rawat, Y. S.; and Shah, M. 2022. Self-supervised learning for videos: A survey. *arXiv preprint arXiv:2207.00419*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *CVPR*, 2818–2826.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. *arXiv preprint arXiv:2203.12602*.

Verma, P.; and Berger, J. 2021. Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions. *arXiv preprint arXiv:2105.00335*.

Wang, R.; Chen, D.; Wu, Z.; Chen, Y.; Dai, X.; Liu, M.; Jiang, Y.-G.; Zhou, L.; and Yuan, L. 2021. Bevt: Bert pretraining of video transformers. *arXiv preprint arXiv:2112.01529*.

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

Xiao, F.; Tighe, J.; and Modolo, D. 2022. MaCLR: Motion-Aware Contrastive Learning of Representations for Videos. In *ECCV*, 353–370. Springer.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 6023–6032.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.