Supplementary Materials

A. Additional Algorithm Details

Algorithm 1 gives an overview of our proposed adversarial mutual information distillation (AMID) framework. The goal of AMID is to learn enhanced target modality representations by distilling the information from multimodal data. AMID simultaneously maximizes the mutual information (MI) between the full modality teacher and the target modality student $I(M;S)$ as well as MI between the teacher and an additional auxiliary modality model $I(M;A)$, while minimizing the conditional entropy of the teacher given the student $H(M|S)$. The maximization of MI is achieved by maximizing its lower bound that takes into account the correlation of samples within a class and an adversarial learning approach is introduced to minimize the conditional entropy.

B. Additional Analysis and Results

B.1. Detailed Implementation

The detailed implementation for knowledge distillation from acoustic to visual modality on UCF51 is presented in this subsection. All experiments are conducted on one 1080Ti. A 1D-CNN14 [24] pretrained on AudioSet [15] is used as the audio extractor, and R(2+1)D-18 [42] is used as the student video network. The fusion module is a 2-layer MLP and the two discriminators are 5-layer MLP and 3-layer MLP respectively. The detailed framework is shown in Fig. 1. Leaky ReLU activation functions are used in all networks. The dimension of representations is 512. The batch size is set to 16 and the temperature τ is set to 0.5. Trainable parameters of the video network and the fusion module are optimized by SGD using the same learning rate of 0.01 with a weight decay of 5e-4 and a momentum of 0.9. Discriminators are trained both with a learning rate of 1e-4. And the initial weights for the two discriminators $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ are set to 0 and 1.0 in all experiments for paying more attention to the alignment among representations of the different samples with the same class of the teacher and the student at the beginning of training. The models are trained for 450 epochs and the learning rates decay 10 times at the 225th, 325th and 400th epoch, respectively. Hyperparameters $\alpha_1, \alpha_2, \alpha_3$ are set to $4, 1.6, 1$, and T_{start} are set to 20.

Table 1. Performance comparison on IEMOCAP, where the target modality is text and the auxiliary modality is video.

Distillation	Method	WA	UA
Video $\rightarrow$ Text	Baseline	68.0	68.7
	GMC	67.6	69.1
	CRCD	68.2	69.2
	CCL	**68.6**	69.2
	AMID	**68.6**	**70.0**

Table 2. More ablative experiments on UCF51.

Configuration	Top-1 accuracy
Baseline	66.7
w/o $\mathcal{L}_{jsd}$	70.4
w/o $\mathcal{L}_{MI_s}$	68.9
w/o $\mathcal{L}_{MI_a}$	70.3
w/o $\mathcal{L}_{adv}$	72.8
AMID	**73.8**

B.2. More Comparison on IEMOCAP

In UCF51 and ActivityNet datasets, we choose video as the target modality because the accompanying audio sometimes is semantically irrelevant to the activities. Audio is better used as the auxiliary modality for the video modality performing recognition and retrieval tasks on the two datasets. In IEMOCAP, all modalities can be used as the target modality to conduct the emotion classification task. To further verify the effectiveness of AMID, we show the results using text as the target and video as the auxiliary one in Tab. 1. Despite video is the weakest modality in IEMOCAP, AMID can still transfer knowledge from video to text and achieve competitive results.

B.3. Ablation Study

To evaluate the contribution of each component in the overall loss function Eq.(16), we conduct more ablations on UCF51 and list the results in Tab. 2. We study the effect of these parts by comparing AMID to four ablative cases: w/o $\mathcal{L}_{jsd}$, w/o $\mathcal{L}_{MI_s}$, w/o $\mathcal{L}_{MI_a}$ and w/o $\mathcal{L}_{adv}$, which remove one constraint at a time. $\mathcal{L}_{jsd}$ and $\mathcal{L}_{MI_s}$ perform the alignment between the teacher and the student jointly in the predictive and latent space. The corresponding results of

Algorithm 1 Adversarial Mutual Information Maximization

Inputs:

$\{x_1, \cdots, x_k\}$: target modality input;
$\{x_{k+1}, \cdots, x_n\}$: auxiliary modality input;
$\{x_1, \cdots, x_n\}$: full modality input;
T, T_{start} and $\mathcal{B}$: the total number of epochs, warm-up period and batch size;
$\tilde{\lambda}(1)$: initial dynamic weights;
$\alpha_1, \alpha_2, \alpha_3$: weights of different loss terms.

Outputs:

M, S, A: the full modality, target modality network and auxiliary modality respectively;
$\mathcal{D}_{\theta_1}$, $\mathcal{D}_{\theta_2}$: the two discriminators.

for $t = 1, ..., T$ **do**
  if $t <= T_{start}$ **then** ▷ warm-up for preliminary alignment
    for every mini-batch $\mathcal{B}$ **do**
      obtain full, target and auxiliary modality representations: $M(x_1, \cdots, x_n)$, $S(x_1, \cdots, x_k)$, $A(x_{k+1}, \cdots, x_n)$
      Update M, S and A using $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{JSD}$
    end for
    $\tilde{\lambda}(t+1) = \tilde{\lambda}(1)$
  else ▷ AMID for cross-modal KD
    for every batch $\mathcal{B}$ **do**
      obtain full, target and auxiliary modality representations: $M(x_1, \cdots, x_n)$, $S(x_1, \cdots, x_k)$, $A(x_{k+1}, \cdots, x_n)$
      Calculate $\mathcal{L}_{MI_s}$ by Eq.(8) // maximize the MI between the teacher and the student.
      Calculate $\mathcal{L}_{MI_a}$ by Eq.(9) // maximize the MI between the teacher and the auxiliary modality.
      Calculate $\mathcal{L}_{adv}$ by Eq.(12) // minimize the conditional entropy of the teacher given the student.
      Update M, S and A using $\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{JSD} + \alpha_1 \mathcal{L}_{MI_s} + \alpha_2 \mathcal{L}_{MI_a} + \alpha_3 \mathcal{L}_{adv}$
      Update D_{θ_1} and D_{θ_2} through minimizing Eq.(13)
    end for
    $\tilde{\lambda}(t+1) = \beta \tilde{\lambda}(t) + (1-\beta)\left(1 - \text{Avg}_t\left(\Phi(M, S)\right)\right)$, $\lambda(t+1) = \sigma(\tilde{\lambda}(t+1))$
    // update dynamic weights through the cosine similarity.
  end if
end for

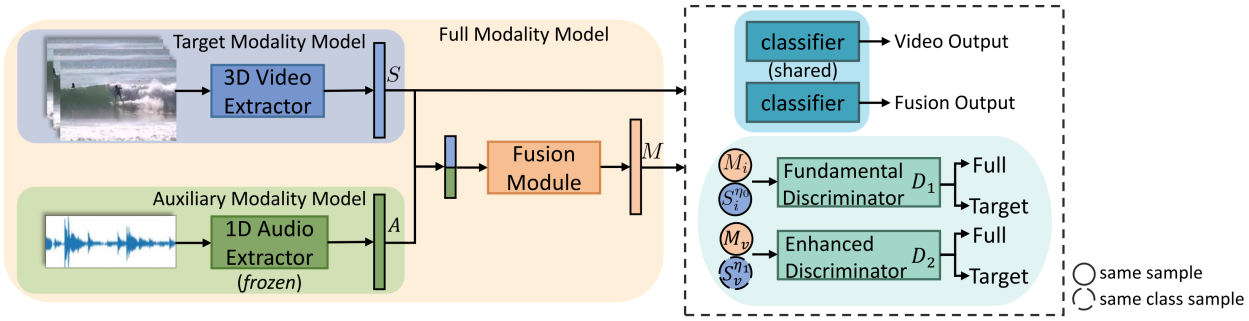

Figure 1. An overview of the proposed AMID for transferring knowledge from acoustic to visual modality on UCF51.

AMID decrease by 3.8% (73.8-70.4) and 4.9% (73.8-68.9) separately by removing one of $\mathcal{L}_{jsd}$ and $\mathcal{L}_{MI_s}$ at a time, which confirms the complementary benefits of them. In addition, the constraining effect of MI on adversarial learning may also be the reason why the performance degrades a lot. Removing $\mathcal{L}_{MI_a}$ decreases the performance by 1.1% (73.8-72.7), indicating its effectiveness in capturing more multimodal information. And the performance degradation resulting from removing $\mathcal{L}_{adv}$ verifies its effect. All of these results indicate that every part in Eq.(17) is necessary and

Table 3. Different strategies of two discriminators on UCF51.

Strategy	Uniform	Dynamic
w/o D_2	72.8	72.8
identical	73.2	\
AMID	72.7	**73.8**

Table 4. Sensitivity Analysis on the smoothing parameter β on UCF51.

β	0	0.5	0.9	0.99	0.999
accuracy	71.0	72.3	73.8	73.4	72.8

works synergistically to distill information across modalities.

We also explore the impact of using different inputs for the two discriminators. We explore this using UCF51, all the configurations are consistent with AMID except the inputs of the discriminators. We conduct experiments with two strategies: (1) two discriminators with identical inputs; (2) all teacher-student pairs are fed into the fundamental discriminator (D_1), while the enhanced discriminator (D_2) deals with pairs from classes occurring with more than once, which corresponds to the strategy used in AMID.

We show results using uniform and dynamic weights for these two strategies in Tab. 3. It can be seen that using one discriminator for training is not sufficient because the performance of strategy (1) with uniform weight is better than that of AMID without D_2, this may be because using one discriminator is not enough to align the representations with the same sample or just same class simultaneously. Besides, using two discriminators in strategy (2) with uniform weights may sometimes be worse than only using a single discriminator, this again demonstrates the necessity to dynamically adjust the weights depending on the alignment of different pairwise representations. In addition, the performance of strategy (1) with uniform weights is worse than that of strategy (2) with dynamic weights, which suggests that paying attention to the pairs coming from the classes with more than one sample can indeed help capture the correlation among samples.

B.4. Parameter Analysis

Moving average. The moving average parameter β, which balances the past alignment with the present states, is applied to improve the training stability. The larger the value is, the smaller the impact of the current alignment has on the dynamic weights. The study on β is presented in Tab. 4. The optimal value of 0.9 is used for all our conducted experiments.

Temperature. The temperature τ is used to adjust the comparability between the positive and negative pairs. Figure 2 reports the results when the τ varies from 0.01 to 1. It

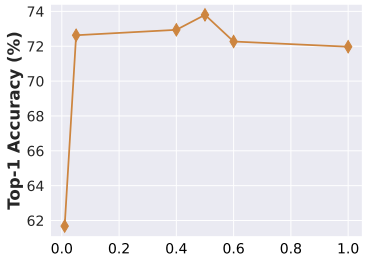

Figure 2. Top-1 accuracy of varying temperature τ on UCF-51.

can be seen that either a very high or low value of τ leads to worse performance. As a result, we set $\tau = 0.5$ for all other experiments.