# Enhanced Audio Tagging via Multi- to Single-Modal Teacher-Student Mutual Learning

**Yifang Yin,[1] Harsh Shrivastava,[1] Ying Zhang*,[1,4]**
**Zhenguang Liu,[2] Rajiv Ratn Shah,[3] Roger Zimmermann[1]**
[1]National University of Singapore, [2]Zhejiang Gongshang University
[3]MIDAS Lab, IIIT-Delhi, [4]Northwestern Polytechnical University, China
{idsyin, dcsrz}@nus.edu.sg, {harsh.vardhan.shri, yingz118, liuzhenguang2008}@gmail.com, rajivratn@iiitd.ac.in

## Abstract

Recognizing ongoing events based on acoustic clues has been a critical yet challenging problem that has attracted significant research attention in recent years. Joint audio-visual analysis can improve the event detection accuracy but may not always be feasible as under many circumstances only audio recordings are available in real-world scenarios. To solve the challenges, we present a novel visual-assisted teacher-student mutual learning framework for robust sound event detection from audio recordings. Our model adopts a multi-modal teacher network based on both acoustic and visual clues, and a single-modal student network based on acoustic clues only. Conventional teacher-student learning performs unsatisfactorily for knowledge transfer from a multi-modality network to a single-modality network. We thus present a mutual learning framework by introducing a single-modal transfer loss and a cross-modal transfer loss to collaboratively learn the audio-visual correlations between the two networks. Our proposed solution takes the advantages of joint audio-visual analysis in training while maximizing the feasibility of the model in use cases. Our extensive experiments on the DCASE17 and the DCASE18 sound event detection datasets show that our proposed method outperforms the state-of-the-art audio tagging approaches.

## 1 Introduction

Audio tagging aims to automatically detect the presence of multiple events in an audio recording, which is of great significance in many applications including surveillance, video indexing, context-aware services, *etc.* (Tian et al. 2018; Imoto et al. 2019; Shrivastava et al. 2020). A precisely recognized sound event can also be used as priors to improve the performance of acoustic scene classification (Wu et al. 2019; Xuan et al. 2020). Existing work mostly leveraged audios as the single modality input, while the attempt of transfer learning in this field is usually limited to fine-tuning the pre-trained acoustic models on large-scale YouTube videos (Hershey et al. 2017).

Combining complementary information from multiple modalities is intuitively appealing for improving the performance of machine learning based approaches (Liu et al. 2018; Yu et al. 2019; Gao et al. 2020). However, the multi-modal features required by such models may not always be available in real-world scenarios. For example, Figure 1 shows an audio surveillance system deployed in domestic environments for anomaly detection such as home abuse. Compared with video cameras, audio sensors (as microphones) have several appealing features including (1) microphones have a 360° coverage whereas standard cameras are constrained by a limited field of view; and (2) Microphones can acquire audio events even when there are obstacles present along the path (Crocco et al. 2016). Thus, it is crucial to develop an effective sound event detection method based on audio recordings when the visual modality is not available due to view obstruction or camera destruction made by people on purpose.

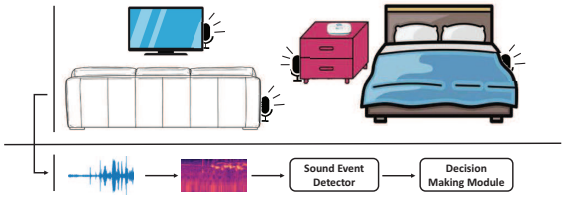


Figure 1: Illustration of an audio surveillance system in domestic environments.

To embrace the advantage and reduce the limitation of using signals from different modalities, we investigate using a multi-modal network to improve the performance of a single-modal network where the latter has no dependence on the auxiliary modalities during inference. To the best of our knowledge, there are currently two popular transfer learning techniques that are widely adopted in previous work in audio tagging: 1) teacher-student network based knowledge transfer and 2) pre-trained model based knowledge transfer. Conventional teacher-student networks are designed for knowledge transfer within a single modality from a large and powerful teacher network to a small student network, which performs less satisfactory for cross-modal knowledge transfer (Hinton, Vinyals, and Dean 2015). Pre-trained model based methods, on the other hand, can learn deep acoustic features from large-scale YouTube videos during training (Aytar, Vondrick, and Torralba 2016). However, as revealed by our experiments, we have found that the pre-

*The corresponding author.

trained acoustic models tend to perform less satisfactorily on audios compared to the pre-trained visual models on images. A key reason might be that audio analysis is more challenging due to the short duration of sound events and the low dimension of the raw waveforms. To fill the research gap, we propose a novel multi- to single-modal teacher-student mutual learning framework. The teacher network performs more robust inference as it fuses features from different modalities, which carry complementary information about different aspects of objects, events, and activities. To conduct effective knowledge transfer, we propose a mutual learning paradigm and introduce two new loss functions to model the audio-visual correlations: 1) a single-modal transfer loss that matches the intermediate acoustic representations, and 2) a cross-modal transfer loss that matches the high-level representations of pairwise audio and visual data. Mutual learning is conducted where both the networks are supervised by the ground-truth labels while matching to the soft predictions and the intermediate representations generated by its peer. We jointly update the parameters of the two networks in each iteration until converge.

Our proposed method is essentially different from the existing transfer learning techniques. Compared to conventional teacher-student networks, the teacher network in our method is powerful in the sense of learning from multiple modalities, rather than having an extremely deep and large architecture. Existing pre-trained models learn a general acoustic representation from large-scale YouTube videos, while our solution learns from relatively small-scale and task-specific videos, which is shown to be more effective. And more importantly, it can be integrated with the general pre-trained models as well. Here we summarize the key contributions of this work as follows:

- We are the first to enhance a single-modal audio-based sound event detector by introducing a visual-assisted multi-modal network that takes video frames as an additional input.
- We present a teacher-student mutual learning framework, which performs effective cross-modal knowledge transfer from the multi-modal teacher network to the single-modal student network.
- We show that our proposed framework is parallel to, and can be easily integrated with existing pre-trained acoustic models, which is a popular transfer learning approach in multimedia analysis.
- We have performed extensive experiments on the DCASE17 and the DCASE18 sound event detection benchmark datasets. Based on audio only, our proposed method outperforms the state-of-the-art audio tagging approaches by a significant margin.

Under many circumstances, the acoustic input and the visual input can be temporally inconsistent. For example, we may hear a dog's barking without actually seeing a dog in the video. We address this problem by modeling the single- and cross-modal transfer losses between the global contexts of the paired modalities. Moreover, the mutual learning we proposed enables the multi-modal teacher network to learn from the single-modal student network as well, which helps reduce overfitting on the visual input to the sound event occurrence, leading to improved audio tagging performance.

## 2 Related Work

The release of the AudioSet (Hershey et al. 2017), which is a large-scale audio dataset annotated with 527 sound event labels, have motivated the development of deep learning approaches in the field of audio analysis (Fayek and Kumar 2020; Wang, Tran, and Feiszli 2020). State-of-the-art classification results have been reported in challenging problems such as acoustic scene classification and sound event detection (Chou, Jang, and Yang 2018; Yin, Shah, and Zimmermann 2018). For example, Kong *et al.* presented a deep CNN with eight layers to detect and localize sound events in audios (Kong et al. 2018). Chou *et al.* proposed an attention-based network architecture that was trained by considering both clip-level and segment-level supervisions (Chou, Jang, and Yang 2018). Yin *et al.* proposed to use a 3D CNN to capture the spatial-temporal dynamic patterns for acoustic scene classification (Yin, Shah, and Zimmermann 2018). Chen *et al.* presented a class-aware self-attention model, which aims at generating discriminative clip-level feature representations for sound event detection (Chen et al. 2018). However, these methods extract features from audio only without exploiting the use of other supplementary data sources.

Transfer learning can be one effective technique that improves the classification performance of one task by transferring knowledge from other data sources (Kumar et al. 2019; Perez et al. 2020). Kumar *et al.* presented a CNN-based framework for sound event detection where adaptation layers were introduced to adapt pre-trained models for new tasks (Kumar, Khadkevich, and Fügen 2017). The representations extracted from the intermediate layers of pre-trained models, such as the SoundNet (Aytar, Vondrick, and Torralba 2016), can be directly used as an additional feature to improve the classification accuracy. Techniques based on teacher-student learning (Li et al. 2017; Tang et al. 2019; Ge et al. 2019) and mutual deep learning (Zhang et al. 2018) have also been proposed recently to transfer knowledge between models. However, the existing efforts have only been made on models with the same single-modal input. A recent trend in multimedia is the joint audio-visual analysis as the two components are naturally correlated in a video (Parekh et al. 2018; Zhou et al. 2018; Hao, Zhang, and Guan 2018). For instance, Owens and Efros proposed to learn a joint representation, which has been shown to be effective for applications including sound source localization and action recognition (Owens and Efros 2018). However, one drawback of such methods is that both the audio and the visual components are required during the inference phase.

## 3 Visual-Assisted Audio Tagging

**Problem.** Let $\mathbf{X}$ denote a set of video samples, and $\mathbf{x} = (\mathbf{a}, \mathbf{v})$ represent the audio and visual components of a video $\mathbf{x} \in \mathbf{X}$. Let $\mathbf{y} \in \{0, 1\}^C$ denote a binary vector of clip-level labels, where $C$ is the number of sound events to be detected. Our goal is to mutually train a teacher network based on

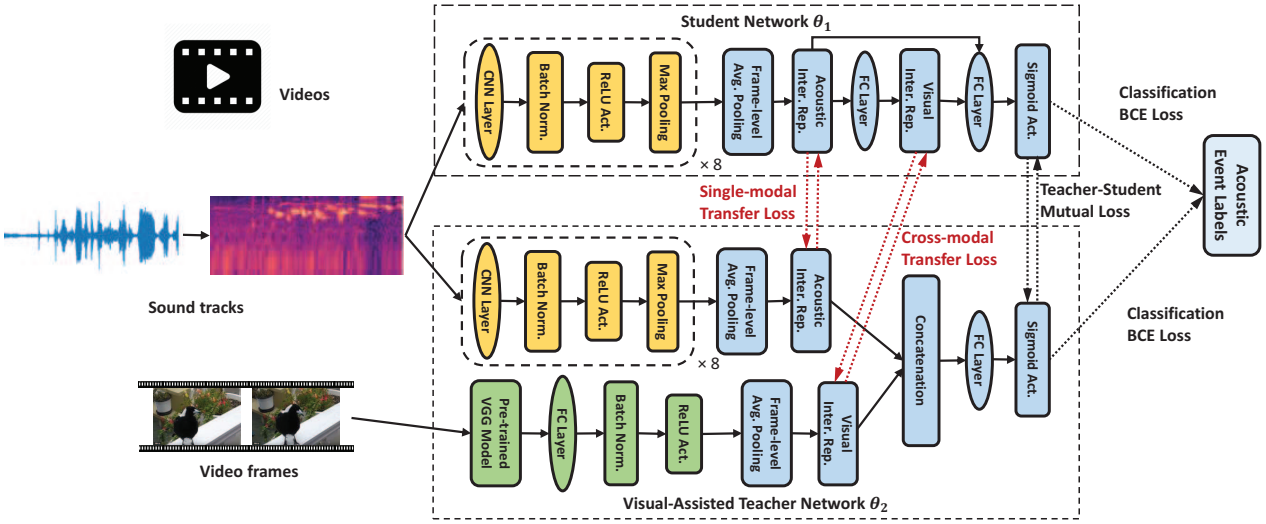


Figure 2: Overview of the proposed teacher-student mutual learning networks.

both the visual and the audio inputs $D = \{(\mathbf{a}, \mathbf{v}, \mathbf{y})|\mathbf{x} \in \mathbf{X}\}$ and an enhanced student network based on only the audio input $D_a = \{(\mathbf{a}, \mathbf{y})|\mathbf{x} \in \mathbf{X}\}$ supervised by clip-level labels $\mathbf{y}$. Next, we introduce the technical details of our proposed teacher-student networks.

## Challenges

Conventional teacher-student learning was proposed for model compression within a single modality, which aimed at training a less expensive student model supervised by an expensive teacher model while maintaining the prediction accuracy (Hinton, Vinyals, and Dean 2015; You et al. 2017). Recently, a few efforts have been made on single-modality domain adaptation (Li et al. 2017; Meng et al. 2018; Fukuda et al. 2017) and feature enhancement (Watanabe et al. 2017) via teacher-student learning. However, both the teacher and the student networks are limited to acoustic models in the aforementioned methods. We thus present the first audio-based sound event detector enhanced by transfer learning from a visual-assisted multi-modal teacher network that takes video frames as an additional input.

## Network Architecture Design

The system overview of our proposed multi- to single-modal teacher-student network is illustrated in Figure 2. We introduce two new losses to model the audio-visual correlations and perform mutual learning to achieve effective cross-modal knowledge transfer.

**Multi-Modal Teacher Network** The teacher network takes both sound tracks and video frames as the input. For audio processing we extract the log-mel spectrograms with 64 bin mel-scale, which are fed to an audio processing sub-network that ranked the third place in the DCASE challenge 2018 (Kong et al. 2018). As shown in Figure 2, the network consists of eight-layer CNNs with the filter size set to $3 \times 3$ and the number of filters set to 64, 64, 128, 128, 256, 256, 512, 512, respectively. Batch normalization (BN) and rectified linear unit (ReLU) are employed after each convolutional layer. Additionally, a max pooling of $2 \times 2$ is applied after layers 2, 4, 6, and 8. To extract visual features, we sample frames at 1 Hz and adopt the pre-trained VGG16 model (Simonyan and Zisserman 2014) to extract a deep feature vector. This deep feature vector is next passed to a fully-connected layer with 128 hidden units, followed by batch normalization and ReLU activation to generate frame-level visual representations. Both the acoustic and the visual frame-level representations are aggregated to clip-level representations based on average pooling, denoted as $\mathbf{h}_\mathbf{t}^\mathbf{a}$ and $\mathbf{h}_\mathbf{t}^\mathbf{v}$, which are next fused by concatenation, denoted as $\mathbf{h_t} = concat(\mathbf{h}_\mathbf{t}^\mathbf{a}, \mathbf{h}_\mathbf{t}^\mathbf{v})$. The feature vector after fusion, $\mathbf{h_t}$, is next fed to a fully-connected layer and Sigmoid activation to output the predictions, termed as $\mathbf{p_t}$, on the sound event classes.

**Single Modal Student Network** The student network takes the sound tracks as the only input. Basically, we adopt the same audio processing sub-network as the one used in the teacher network (Kong et al. 2018), but with the following modifications to support effective cross-modal knowledge transfer. It has been shown in previous literatures that using the intermediate representations learnt by the teacher network to supervise the training of the student network can achieve potential performance gain (Romero et al. 2014). Let $\mathbf{h}_\mathbf{s}^\mathbf{a}$ denote the acoustic intermediate representations generated by the student network. We model the single-modal transfer loss within the acoustic modality as,

$$L_{Single} = \sum_{i=1}^{N} \|\mathbf{h}_\mathbf{t}^\mathbf{a}(a_i) - \mathbf{h}_\mathbf{s}^\mathbf{a}(a_i)\|^2 \quad (1)$$

where $N$ is the number of samples in the training dataset. As this single-modal transfer loss can only propagate knowledge in the acoustic modality from the teacher network to the student network, we need to model the cross-modal cor-

relations between the audio and video frames to further achieve cross-modal knowledge transfer. The sound tracks and video frames are naturally correlated in each video. Though the segment-level acoustic and visual features can be unsynchronized due to temporal inconsistency, the global contexts of different modalities are mostly correlated as they tend to share high-level concepts such as the clip-level events (Xuan et al. 2020). Here we use the average pooling operation to summarize the representations of the input sequence to obtain the global context of the acoustic modality $\mathbf{h_s^a}$ in the student network and the global context of the visual modality $\mathbf{h_t^v}$ in the teacher network. The main idea of our cross-modal knowledge transfer is to let the global context representations of the paired modalities similar to each other. Thus we model the cross-modal transfer loss as,

$$L_{Cross} = \sum_{i=1}^{N} \|\mathbf{h_t^v}(v_i) - \phi(\mathbf{W_c h_s^a}(a_i) + \mathbf{b_c})\|^2 \quad (2)$$

where $\mathbf{W_c}$ and $\mathbf{b_c}$ are the trainable parameters that match the size of acoustic global context in the student network to the size of the visual global context in the teacher network, $\phi$ is an activation function where ReLU is adopted in our framework. Similar to the teacher network, we fuse $\mathbf{h_s^a}(a_i)$ and $\phi(\mathbf{W_c h_s^a}(a_i)+\mathbf{b_c})$ and pass it to a fully-connected layer with Sigmoid activation to output the predictions, termed as $\mathbf{p_s}$, on the sound event classes.

**Multi- to Single-Modal Mutual Learning** To achieve effective knowledge transfer from multi-modal to single-modal network, we further present a teacher-student mutual learning framework where we formulate audio tagging as a multi-class classification problem as multiple sound events may occur in a single audio clip. The supervised classification loss is computed using the binary cross-entropy error between the student predictions, teacher predictions, and ground-truth labels as,

$$\begin{aligned} L_{Class}^{student} &= BCE(\mathbf{y}, \mathbf{p_s}) + \lambda_s BCE(\mathbf{p_t}, \mathbf{p_s}) \\ &= -\sum_{i=1}^{N}\sum_{k=1}^{K}(y_i^k \log(p_s^k(a_i)) + (1-y_i^k)\log(1-p_s^k(a_i))) \\ &- \lambda_s \sum_{i=1}^{N}\sum_{k=1}^{K}(p_t^k(x_i) log(p_s^k(a_i)) + (1-p_t^k(x_i))\log(1-p_s^k(a_i))) \end{aligned} \quad (3)$$

$$\begin{aligned} L_{Class}^{teacher} &= BCE(\mathbf{y}, \mathbf{p_t}) + \lambda_t BCE(\mathbf{p_s}, \mathbf{p_t}) \\ &= -\sum_{i=1}^{N}\sum_{k=1}^{K}(y_i^k \log(p_t^k(x_i)) + (1-y_i^k)\log(1-p_t^k(x_i))) \\ &- \lambda_t \sum_{i=1}^{N}\sum_{k=1}^{K}(p_s^k(a_i) log(p_t^k(x_i)) + (1-p_s^k(a_i))\log(1-p_t^k(x_i))) \end{aligned} \quad (4)$$

where $L_{Class}^{teacher}$ and $L_{Class}^{student}$ are the supervised classification losses for the teacher network and the student network, respectively. $y_i^k$ represents the ground-truth label of the $i$-th sample for class $k$. Recall that $\mathbf{x} = (\mathbf{a}, \mathbf{v})$ and $\mathbf{a}$ denote the multi-modal and single-modal inputs to the teacher and the student networks, thus $p_t^k(x_i)$ and $p_s^k(a_i)$ represent the predicted scores of the $i$-th sample for class $k$ generated by the teacher network and the student network, respectively.

Finally, by fusing the supervised classification loss with our proposed single-modal transfer loss and cross-modal transfer loss, we obtain the overall losses for the two networks as

$$L_{student} = L_{Class}^{student} + \gamma_s L_{Single} + \omega_s L_{Cross} \quad (5)$$

$$L_{teacher} = L_{Class}^{teacher} + \gamma_t L_{Single} + \omega_t L_{Cross} \quad (6)$$

where $\lambda$, $\gamma$, and $\omega$ are balancing factors that control the weights of different losses. In our framework, not only the student network learns from its teacher, the teacher network also adapts its "teaching strategy" to its student. It is worth mentioning that such a collaborative training strategy can help align the global contexts in different modalities much better in the feature space compared to the conventional one-way training. The two models are optimized jointly in every mini-batch. At each iteration, we compute the intermediate representations based on the current models, and update both networks' parameters according to the intermediate representations of its peer, the soft outputs of its peer, and the ground-truth labels of training samples. We collaboratively train the two models until converge.

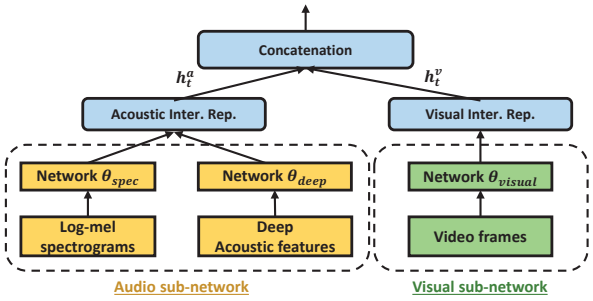


Figure 3: Illustration of a generalized teacher network that fuses multiple acoustic features extracted from the input.

## Model Generalization and Multi-Feature Fusion

We have introduced how to train a student network based on a single acoustic feature, *i.e.*, the log-mel spectrograms. Here we generalize our model by fusing multiple acoustic features in both the student and the teacher networks to further boost the system's performance. Figure 3 shows a generalized architecture of the teacher network. We process each of the input acoustic features such as log-mel spectrograms and pre-trained deep acoustic representations (Hershey et al. 2017) by separate feature-processing sub-networks to generate the global context from each acoustic feature. The global contexts are next fused by concatenation to form the acoustic intermediate representation $\mathbf{h_a^t}$. The same audio sub-network can be adopted in the corresponding student model to generate $\mathbf{h_a^s}$. Thus, we can see that the multi-feature fusion can be easily integrated in our proposed framework to further improve the audio tagging results.

# 4 Evaluation

We first introduce the experimental setup, and then proceed with the evaluations by performing a step-by-step model justification and a comparison with the state-of-the-art audio tagging approaches.

Table 1: Comparison of audio tagging based on audio only, visual only, and their fusion.

(a) DCASE17

| Feature | Source | mAP | mAUC |
|---|---|---|---|
| Log-mel spectrogram | Audio | 0.5918 | 0.9162 |
| VGG - 1000 | Visual | 0.4512 | 0.8525 |
| VGG - 4096 | Visual | 0.5225 | 0.8752 |
| Log-mel spec. + VGG-1000 | Fusion | 0.6763 | **0.9396** |
| Log-mel spec. + VGG-4096 | Fusion | **0.6824** | 0.9375 |

(b) DCASE18

| Feature | Source | mAP | mAUC |
|---|---|---|---|
| Log-mel spectrogram | Audio | 0.8637 | 0.9512 |
| VGG - 1000 | Visual | 0.6930 | 0.8991 |
| VGG - 4096 | Visual | 0.7750 | 0.9298 |
| Log-mel spec. + VGG-1000 | Fusion | 0.8818 | 0.9617 |
| Log-mel spec. + VGG-4096 | Fusion | **0.9222** | **0.9773** |

Table 2: Mean average precision comparison of the student network trained using different strategies.

(a) DCASE17

| Training Strategy | mAP | mAUC | mAP Decrease |
|---|---|---|---|
| Proposed mutual learning | 0.6421 | 0.9292 | – |
| - w/o mutual learning | 0.6359 | 0.9239 | -1.0% |
| - w/o single-modal transfer loss | 0.6317 | 0.9234 | -1.6% |
| - w/o cross-modal transfer loss | 0.6353 | 0.9280 | -1.1% |

(b) DCASE18

| Training Strategy | mAP | mAUC | mAP Decrease |
|---|---|---|---|
| Proposed mutual learning | 0.8855 | 0.9645 | – |
| - w/o mutual learning | 0.8760 | 0.9599 | -1.1% |
| - w/o single-modal transfer loss | 0.8702 | 0.9625 | -1.7% |
| - w/o cross-modal transfer loss | 0.8775 | 0.9618 | -0.9% |

## Experimental Setup

We evaluated our proposed method on two public audio datasets, namely the DCASE17 sound event detection for smart cars and the DCASE18 sound event detection in domestic environments. Both datasets are subsets of the AudioSet (Hershey et al. 2017). The DCASE17 dataset contains more than 50K audio clips annotated with 17 sound events of smart cars, while the DCASE18 dataset focuses on 10 classes of sound events in domestic environments. All the samples are 10-second sound clips drawn from YouTube videos. Following previous work, we use mean average precision (mAP), and mean area under ROC curve (mAUC) as the evaluation metrics (Yin et al. 2019).

The balancing factors $\lambda$, $\gamma$, and $\omega$ in the loss functions are empirically set to 1 throughout the experiments. For optimization, we train the neural networks using the Adam optimizer with a batch size of 32. The learning rate is set to 0.001. To reduce the impact of randomness in neural networks. We train each of the models three times with random seeds and take their average as the final prediction scores as the final results (Krogh and Vedelsby 1994).

## Model Justification

We perform a step-by-step model justification to demonstrate our choice of the input visual features and the effectiveness of our proposed multi- to single-modal mutual learning solution, both of which are the key components in our audio tagging framework.

**Evaluation on features** In terms of the visual modality, we investigate two types of deep features generated by the pre-trained VGG16 model on the ImageNet (Simonyan and Zisserman 2014). VGG-1000 represents the 1000-dimensional feature vector generated by the output layer, while VGG-4096 represents the 4096-dimensional feature vector generated by the last layer before the output layer. Though the use of VGG-1000 has been investigated by Aytar *et al.*, which obtained promising results on acoustic scene classification, VGG-4096 turned out to be more effective for audio tagging. This is based on the observation that VGG-4096 outperformed VGG-1000 on both DCASE17 and DCASE18 datasets by 15.8% and 11.8% in terms of the mAP, respectively. By fusing the log-mel spectrograms and the VGG-4096 deep feature, the mAP and the mAUC have been improved by 15.3% and 2.3% on the DCASE17 dataset and by 6.8% and 2.7% on the DCASE18 dataset, respectively, compared with the audio baseline method using the log-mel spectrograms only.

**Evaluation on multi- to single-modal mutual learning** Our proposed visual-assisted audio tagging framework consists of three main components to perform effective knowledge transfer from multi-modal to single-modal networks. To demonstrate the effectiveness of our proposed method in each step, we train the networks by removing one of the components at each time (*i.e.*, mutual learning, single-modal transfer loss, and cross-modal transfer loss) and report the results in Table 2. As can be seen, without mutual learning, the mAP obtained by the student network decreased by 1.0% and 1.1% respectively on the two datasets. This indicates that a student is able to learn more effectively from a teacher, which adapts its supervision according to the student at the same time. The single-modal transfer loss and the cross-modal transfer loss address the temporal inconsistency between audio and video by aligning the global contexts of different modalities in the high-level feature space. Without them in the overall losses, the mAP obtained by the student network can decrease as much as 1.7%, thus verifying the effectiveness of our proposed approach.

## Comparison with the State-of-the-art

We compare our method to both the state-of-the-art transfer learning methods and the state-of-the-art audio tagging methods to demonstrate the effectiveness of our proposed approach. The Sup. Data column in Tables 3, 4, and 5 indicates if a supplementary large-scale YouTube video dataset

Table 3: Comparison to the state-of-the-art teacher-student network based knowledge transfer.

| Method | Sup. Data | DCASE17 | | DCASE18 | |
|---|---|---|---|---|---|
| | | mAP | mAUC | mAP | mAUC |
| Standard Back-propagation | ✗ | 0.5918 | 0.9162 | 0.8637 | 0.9512 |
| Knowledge Distillation (Hinton, Vinyals, and Dean 2015) | ✗ | 0.6138 | 0.9203 | 0.8642 | 0.9563 |
| Enhanced Feature (Watanabe et al. 2017) | ✗ | 0.5961 | 0.9144 | 0.8619 | 0.9549 |
| FitNets (Romero et al. 2014) | ✗ | 0.6250 | 0.9184 | 0.8639 | 0.9583 |
| Mutual Learning (Zhang et al. 2018) | ✗ | 0.6229 | 0.9217 | 0.8633 | 0.9566 |
| Proposed - Student | ✗ | **0.6421** | **0.9292** | **0.8855** | **0.9645** |

Table 4: Comparison to the state-of-the-art pre-trained model based deep acoustic representations learnt from videos.

| Feature | Sup. Data | DCASE17 | | DCASE18 | |
|---|---|---|---|---|---|
| | | mAP | mAUC | mAP | mAUC |
| [1]Log-mel spectrogram (Kong et al. 2018) | ✗ | 0.5918 | 0.9162 | 0.8637 | 0.9512 |
| [2]SoundNet (Aytar, Vondrick, and Torralba 2016) | ✓ | 0.4256 | 0.8581 | 0.7426 | 0.9168 |
| [3]VGGish (Hershey et al. 2017) | ✓ | 0.5786 | 0.9179 | 0.8408 | 0.9582 |
| [4]Log-mel spec. + SoundNet | ✓ | 0.6149 | 0.9216 | 0.8633 | 0.9553 |
| [5]Log-mel spec. + VGGish | ✓ | **0.6480** | 0.9278 | 0.8824 | **0.9678** |
| Proposed - Student | ✗ | 0.6421 | **0.9292** | **0.8855** | 0.9645 |

Table 5: Integration of the deep acoustic representations in our cross-modal teacher-student mutual learning network.

| Proposed - Student | Sup. Data | DCASE17 | | | DCASE18 | | |
|---|---|---|---|---|---|---|---|
| | | mAP | mAUC | mAP Gain | mAP | mAUC | mAP Gain |
| [1]Log-mel spectrogram + (VGG-4096) | ✗ | 0.6421 | 0.9292 | 8.5% | 0.8855 | 0.9645 | 2.5% |
| [2]SoundNet + (VGG-4096) | ✓ | 0.4419 | 0.8709 | 3.8% | 0.7676 | 0.9295 | 3.4% |
| [3]VGGish + (VGG-4096) | ✓ | 0.5965 | 0.9248 | 3.1% | 0.8580 | 0.9635 | 2.0% |
| [4]Log-mel spec. + SoundNet + (VGG-4096) | ✓ | 0.6366 | 0.9269 | 3.5% | 0.8845 | 0.9626 | 2.5% |
| [5]Log-mel spec. + VGGish + (VGG-4096) | ✓ | **0.6690** | **0.9353** | 3.2% | **0.9059** | **0.9704** | 2.7% |

is required in addition to the task-specific training dataset. It is also worth mentioning that we have applied network ensembles (Krogh and Vedelsby 1994) to reduce the randomness in all the methods for a fair comparison. We train each of the models three times with random seeds and take their average as the final predictions. Thus, the result difference shown in the tables can be considered as significant in our experiments.

**Comparison to teacher-student network based transfer learning** Table 3 shows the comparison of our proposed method to the state-of-the-art teacher-student network based transfer learning techniques. Watanabe *et al.* proposed to train a teacher network with enhanced features and use the soft targets of the teacher network to train the student network. Hinton *et al.* proposed to combine the losses to both the ground-truth labels and the soft labels of the teacher network to train the student network. In our experiments, we set the two losses with equal weights. Additionally, as the audio tagging is a multi-label classification, we use the binary cross entropy instead of the KL divergence as the loss function between the student and the teacher predictions. FitNets and Mutual Learning improved the knowledge distillation proposed by Hinton *et al.* in different ways. FitNets proposed to use the intermediate representations learned by the teacher as hints to improve the training process and final performance of the student. In our experiments, we used the acoustic intermediate representation in the student network as the guided layer and the concatenation of the acoustic and the visual intermediate representations in the teacher network as the hint layer (see Figure 2). A fully-connected regressor was adopted to model the loss between the hint and the guided layers. The Mutual Learning method, on the other hand, proposed to train the teacher and the student networks collaboratively instead of a one-way knowledge transfer (Zhang et al. 2018). However, as all the previous teacher-student learning techniques are designed for knowledge transfer within a single modality, they performed less satisfactorily for cross-modal knowledge transfer. Our proposed method outperformed the previous teacher-student networks and obtained the best mAP of 0.6421 and 0.8855 on the DCASE17 and DCASE18 datasets, respectively.

**Comparison to pre-trained model based transfer learning** Table 4 shows the comparison of our proposed method to the state-of-the-art pre-trained model based transfer learning approaches. SoundNet (Aytar, Vondrick, and Torralba 2016) and VGGish (Hershey et al. 2017) are two of the most advanced models that generate deep acoustic features learnt from large-scale YouTube videos. The SoundNet representation was trained on two million unlabeled videos supervised by the semantics extracted from the vision. While the

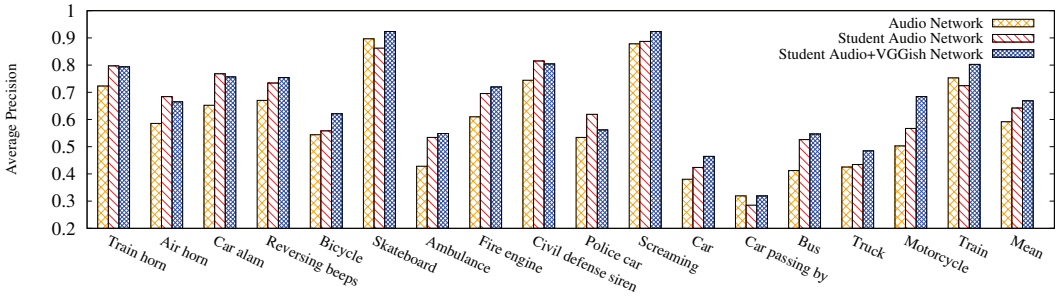


Figure 4: Average precision comparison on the 17 sound event classes of DCASE17 obtained by the conventional audio network, and our proposed student audio network trained with or without the pre-trained VGGish representation.

Table 6: Audio tagging performance comparison (in %) on the DCASE17 test dataset.

| Methods | Precision | Recall | F1 |
|---|---|---|---|
| (Lee, Park, and Nam 2017) | 37.6 | 45.7 | 41.2 |
| (Adavanne and Virtanen 2017) | 47.5 | 39.6 | 43.2 |
| (Salamon, McFee, and Li 2017) | 44.7 | 47.0 | 45.9 |
| (Vu, Dang, and Wang 2017) | 54.2 | 49.5 | 51.8 |
| (Xu et al. 2018) | 53.8 | 60.1 | 56.7 |
| (Lee et al. 2017) | **70.3** | 47.9 | 57.0 |
| (Yin et al. 2019) | 56.0 | 60.6 | 58.2 |
| Proposed - Student | 57.7 | **64.2** | **60.8** |

VGGish representation was trained on a preliminary version of YouTube-8M supervised by machine-generated labels derived from a combination of metadata (title, description, comments, *etc.*), context, and vision. The VGGish representation outperformed the SoundNet representation as the former was trained based on more accurate and diverse labels. However, the performance gain tends to be less satisfactory when considering the large number of supplementary videos required for training. Comparatively, our proposed student network obtained a competitive mAP of 0.6421 and 0.8855 on the two datasets based on log-mel spectrograms only without utilizing any supplementary YouTube datasets during training. Moreover, as introduced in the model generalization section, our method is parallel to the pre-trained model based transfer learning where the learnt deep acoustic representations can be easily integrated into our framework. Next, we evaluate the effectiveness of the generalization of our proposed model.

Table 5 reports the results obtained by the student network after integrating the deep acoustic representations in our proposed framework. The mAP Gain is computed *w.r.t.* the corresponding result in Table 4 marked with the same superscript. Please note that the (VGG-4096) visual feature was only used during training. The paired methods in Tables 4 and 5 used the same acoustic features during inference. As can be seen, our proposed method consistently improved the mAP by 2.0% $\sim$ 8.5% when performing sound event detection based on the same acoustic features, compared to the previous methods. This indicates the effectiveness of our proposed multi- to single-modal teacher-student mutual learning and its generalization.

**Comparison to the top audio tagging systems in the DCASE challenge** Next, we compare our best model to the state-of-the-art, top-ranked systems in the DCASE challenge. We use a global confidence threshold of 0.25 for all the 17 sound events and report the results in Table 6. Lee *et al.* proposed to use an ensemble of convolutional neural networks to detect the weakly labeled audio events. Xu *et al.* presented a gated convolutional neural network with attention-based temporal aggregation method for audio event detection. Yin *et al.* further improved Xu's gated CNN by applying multi-level feature fusion. Our method obtained the best recall and F1 score on the DCASE17 test set. We were able to improve the F1 score by 4.5%, compared to the F1 score of 0.582 obtained by the second best method. Finally, Figure 4 shows the comparison of the per-class average precision. Audio Network refers to the Log-mel spectrogram in Table 4. Student Audio Network and Student Audio+VGGish Network refer to the Log-mel spectrogram and the Log-mel spec.+VGGish in Table 5, which are our proposed student networks. Generally speaking, our proposed solution performs consistently well among different sound events. Student Audio Network and Student Audio+VGGish Network outperformed the conventional Audio Network by 8.5% and 13.0% in terms of mAP, respectively. The improvement is significant as all the three networks perform audio-based sound event detection without any additional input from other modalities during the inference phase.

## 5 Conclusions

We investigate the use of video frames that are associated with the sound tracks in the problem of audio tagging. Different from the existing visual-audio joint analysis methods, our goal is to leverage video frames only in the training phase to improve the accuracy of audio-based sound event detection. We propose a novel teacher-student mutual learning framework to effectively transfer the knowledge of a multi-modal teacher network to a single-modal student network. Experiments on the DCASE17 and the DCASE18 sound event detection datasets showed that our proposed method outperformed the conventional multi-label learning by 8.5% and 2.5% based on log-mel spectrograms and by 13.0% and 4.9% based on a fusion of log-mel spectrograms and Google's VGGish acoustic representations, in terms of the mean average precision.

## 6 Acknowledgments

This research is supported by Singapore Ministry of Education Academic Research Fund Tier 1 under MOE's official grant number T1 251RES1820. Rajiv Ratn Shah is partly supported by the Infosys Center for AI at IIIT Delhi. We also thank ClickUp for providing the Unlimited Plan of its platform to our lab for efficient task management.

## References

Adavanne, S.; and Virtanen, T. 2017. Sound Event Detection Using Weakly Labeled Dataset with Stacked Convolutional and Recurrent Neural Network. *DCASE2017 Challenge* .

Aytar, Y.; Vondrick, C.; and Torralba, A. 2016. SoundNet: Learning Sound Representations from Unlabeled Video. In *International Conference on Neural Information Processing Systems*, 892–900.

Chen, S.; Chen, J.; Jin, Q.; and Hauptmann, A. 2018. Class-aware Self-Attention for Audio Event Recognition. In *ACM International Conference on Multimedia Retrieval*, 28–36.

Chou, S.-Y.; Jang, J.-S.; and Yang, Y.-H. 2018. Learning to Recognize Transient Sound Events using Attentional Supervision. In *International Joint Conference on Artificial Intelligence*, 3336–3342.

Crocco, M.; Cristani, M.; Trucco, A.; and Murino, V. 2016. Audio Surveillance: A Systematic Review. *ACM Computing Surveys* 48(4): 52:1–52:46.

Fayek, H. M.; and Kumar, A. 2020. Large Scale Audiovisual Learning of Sounds with Weakly Labeled Data. In *International Joint Conference on Artificial Intelligence*, 558–565.

Fukuda, T.; Suzuki, M.; Kurata, G.; Thomas, S.; Cui, J.; and Ramabhadran, B. 2017. Efficient Knowledge Distillation from an Ensemble of Teachers. In *Interspeech*, 3697–3701.

Gao, R.; Oh, T. H.; Grauman, K.; and Torresani, L. 2020. Listen to Look: Action Recognition by Previewing Audio. In *IEEE CVPR*, 10454–10464.

Ge, S.; Zhao, S.; Li, C.; and Li, J. 2019. Low-Resolution Face Recognition in the Wild via Selective Knowledge Distillation. *IEEE Transactions on Image Processing* 28(4): 2051–2062.

Hao, W.; Zhang, Z.; and Guan, H. 2018. CMCGAN: A Uniform Framework for Cross-modal Visual-Audio Mutual Generation. In *AAAI Conference on Artificial Intelligence*.

Hershey, S.; Chaudhuri, S.; Ellis, D. P. W.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; Slaney, M.; Weiss, R. J.; and Wilson, K. 2017. CNN Architectures for Large-scale Audio Classification. In *IEEE ICASSP*, 131–135.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* URL https://arxiv.org/abs/1503.02531.

Imoto, K.; Niitsuma, M.; Yamanishi, R.; and Yamashita, Y. 2019. Joint Analysis of Acoustic Event and Scene Based on Multitask Learning. *arXiv preprint arXiv:1904.12146* .

Kong, Q.; Iqbal, T.; Xu, Y.; Wang, W.; and Plumbley, M. D. 2018. DCASE 2018 Challenge Surrey Cross-Task Convolutional Neural Network Baseline. *arXiv preprint arXiv:1808.00773v4* URL https://arxiv.org/abs/1808.00773v4.

Krogh, A.; and Vedelsby, J. 1994. Neural Network Ensembles, Cross Validation and Active Learning. In *International Conference on Neural Information Processing Systems*, 231–238.

Kumar, A.; Khadkevich, M.; and Fügen, C. 2017. Knowledge Transfer from Weakly Labeled Audio using Convolutional Neural Network for Sound Events and Scenes. *arXiv preprint arXiv:1711.01369* URL http://arxiv.org/abs/1711.01369.

Kumar, A.; Shah, A.; Hauptmann, A.; and Raj, B. 2019. Learning Sound Events from Webly Labeled Data. In *International Joint Conference on Artificial Intelligence*, 2772–2778.

Lee, D.; Lee, S.; Han, Y.; and Lee, K. 2017. Ensemble of Convolutional Neural Networks for Weakly-Supervised Sound Event Detection Using Multiple Scale Input. *DCASE2017 Challenge* .

Lee, J.; Park, J.; and Nam, J. 2017. Combining Multi-Scale Features Using Sample-Level Deep Convolutional Neural Networks for Weakly Supervised Sound Event Detection. *DCASE2017 Challenge* .

Li, J.; Seltzer, M. L.; Wang, X.; Zhao, R.; and Gong, Y. 2017. Large-scale Domain Adaptation via Teacher-Student Learning. *arXiv preprint arXiv:1708.05466* URL https://arxiv.org/abs/1708.05466.

Liu, K.; Li, Y.; Xu, N.; and Natarajan, P. 2018. Learn to Combine Modalities in Multimodal Deep Learning. *arXiv preprint arXiv:1805.11730* .

Meng, Z.; Li, J.; Gong, Y.; and Juang, B.-H. 2018. Adversarial Teacher-Student Learning for Unsupervised Domain Adaptation. In *IEEE ICASSP*, 5949–5953.

Owens, A.; and Efros, A. A. 2018. Audio-Visual Scene Analysis with Self-supervised Multisensory Features. In *ECCV*, 631–648.

Parekh, S.; Essid, S.; Ozerov, A.; Duong, N. Q.; Pérez, P.; and Richard, G. 2018. Weakly Supervised Representation Learning for Unsynchronized Audio-Visual Events. *arXiv preprint arXiv:1804.07345* URL https://arxiv.org/abs/1804.07345.

Perez, A.; Sanguineti, V.; Morerio, P.; and Murino, V. 2020. Audio-visual Model Distillation using Acoustic Images. In *IEEE WACV*, 2854–2863.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for Thin Deep Nets. *arXiv preprint arXiv:1412.6550* .

Salamon, J.; McFee, B.; and Li, P. 2017. DCASE 2017 Submission: Multiple Instance Learning for Sound Event Detection. *DCASE2017 Challenge* .

Shrivastava, H.; Yin, Y.; Shah, R. R.; and Zimmermann, R. 2020. MT-GCN for Multi-label Audio-tagging with Noisy Labels. In *IEEE ICASSP*, 136–140.

Simonyan, K.; and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint arXiv:1409.1556* .

Tang, Y.; Lu, J.; Wang, Z.; Yang, M.; and Zhou, J. 2019. Learning Semantics-Preserving Attention and Contextual Interaction for Group Activity Recognition. *IEEE Transactions on Image Processing* 28(10): 4997–5012.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual Event Localization in Unconstrained Videos. In *ECCV*, 247–263.

Vu, T.; Dang, A.; and Wang, J.-C. 2017. Deep Learning for DCASE2017 Challenge. *DCASE2017 Challenge* .

Wang, W.; Tran, D.; and Feiszli, M. 2020. What Makes Training Multi-Modal Classification Networks Hard? In *IEEE CVPR*, 12695–12705.

Watanabe, S.; Hori, T.; Le Roux, J.; and Hershey, J. R. 2017. Student-teacher Network Learning with Enhanced Features. In *IEEE ICASSP*, 5275–5279.

Wu, Y.; Zhu, L.; Yan, Y.; and Yang, Y. 2019. Dual Attention Matching for Audio-visual Event Localization. In *ICCV*, 6292–6300.

Xu, Y.; Kong, Q.; Wang, W.; and Plumbley, M. D. 2018. Large-Scale Weakly Supervised Audio Classification Using Gated Convolutional Neural Network. In *IEEE ICASSP*, 121–125.

Xuan, H.; Zhang, Z.; Chen, S.; Yang, J.; and Yan, Y. 2020. Cross-Modal Attention Network for Temporal Inconsistent Audio-Visual Event Localization. In *AAAI Conference on Artificial Intelligence*.

Yin, Y.; Chiou, M.-J.; Liu, Z.; Shrivastava, H.; Shah, R. R.; and Zimmermann, R. 2019. Multi-Level Fusion Based Class-Aware Attention Model for Weakly Labeled Audio Tagging. In *ACM International Conference on Multimedia*, 1304—1312.

Yin, Y.; Shah, R. R.; and Zimmermann, R. 2018. Learning and Fusing Multimodal Deep Features for Acoustic Scene Categorization. In *ACM International Conference on Multimedia*, 1892–1900.

You, S.; Xu, C.; Xu, C.; and Tao, D. 2017. Learning from Multiple Teacher Networks. In *ACM SIGKDD*, 1285–1294.

Yu, Y.; Tang, S.; Raposo, F.; and Chen, L. 2019. Deep Cross-Modal Correlation Learning for Audio and Lyrics in Music Retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15(1).

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep Mutual Learning. In *IEEE CVPR*, 4320–4328.

Zhou, Y.; Wang, Z.; Fang, C.; Bui, T.; and Berg, T. L. 2018. Visual to Sound: Generating Natural Sound for Videos in the Wild. In *IEEE CVPR*, 3550–3558.