# Multimodal Knowledge Expansion

Zihui Xue[1,2], Sucheng Ren[1,3], Zhengqi Gao[1,4], and Hang Zhao [*5,1]

[1]Shanghai Qi Zhi Institute, [2]UT Austin
[3]South China University of Technology
[4]MIT, [5]Tsinghua University

## Abstract

*The popularity of multimodal sensors and the accessibility of the Internet have brought us a massive amount of unlabeled multimodal data. Since existing datasets and well-trained models are primarily unimodal, the modality gap between a unimodal network and unlabeled multimodal data poses an interesting problem: how to transfer a pre-trained unimodal network to perform the same task with extra unlabeled multimodal data? In this work, we propose multimodal knowledge expansion (MKE), a knowledge distillation-based framework to effectively utilize multimodal data without requiring labels. Opposite to traditional knowledge distillation, where the student is designed to be lightweight and inferior to the teacher, we observe that a multimodal student model consistently rectifies pseudo labels and generalizes better than its teacher. Extensive experiments on four tasks and different modalities verify this finding. Furthermore, we connect the mechanism of MKE to semi-supervised learning and offer both empirical and theoretical explanations to understand the expansion capability of a multimodal student.* [1]

## 1. Introduction

Deep neural networks and supervised learning have made outstanding achievements in fields like computer vision [16, 21, 33] and computer audition [17, 47]. With the popularity of multimodal data collection devices (*e.g.*, RGB-D cameras and video cameras) and the accessibility of the Internet, a large amount of unlabeled multimodal data has become available. A couple of examples are shown in Figure 1: (a) A unimodal dataset has been previously annotated for the data collected by an old robot; after a hardware upgrade with an additional sensor, the roboticist has access to some new unlabeled multimodal data. (b) Internet videos are abundant and easily accessible. While there are existing unimodal datasets and models for tasks such as image recognition, we further want to perform the same task on unlabeled videos. A natural question arises: *how to transfer a unimodal network to the unlabeled multimodal data?*

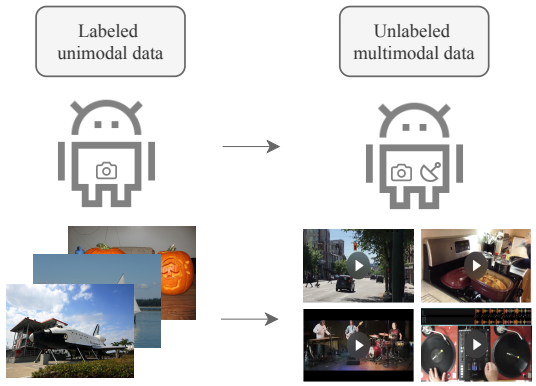



Figure 1: The popularity of multimodal data collection devices and the Internet engenders a large amount of unlabeled multimodal data. We show two examples above: (a) after a hardware upgrade, lots of unannotated multimodal data are collected by the new sensor suite; (b) large-scale unlabeled videos can be easily obtained from the Internet.

One naive solution is to directly apply the unimodal network for inference using the corresponding modality of unlabeled data. However, it overlooks information described by the other modalities. While learning with multimodal data has the advantage of facilitating information fusion and inducing more robust models compared with only using one modality, developing a multimodal network with supervised learning requires tremendous human labeling efforts.

In this work, we propose multimodal knowledge expansion (*MKE*), a knowledge distillation-based framework, to make the best use of unlabeled multimodal data. *MKE* enables a multimodal network to learn on the unlabeled data with minimum human labor (*i.e.*, no annotation of the mul-

*Corresponding to hangzhao@mail.tsinghua.edu.cn
[1]Code is available at: https://github.com/zihuixue/MKE

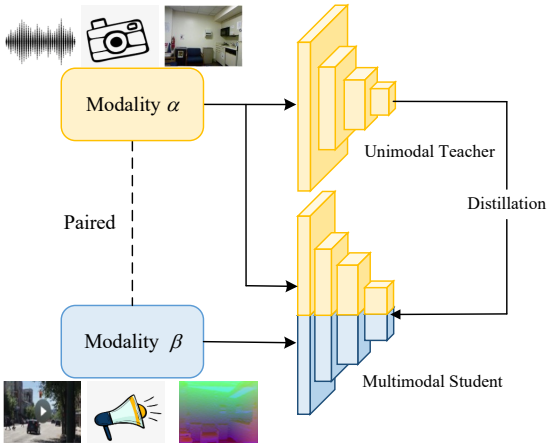



Figure 2: Framework of *MKE*. In *knowledge distillation*, a cumbersome teacher network is considered as the upper bound of a lightweight student network. Contradictory to that, we introduce a unimodal teacher and a multimodal student. The multimodal student achieves *knowledge expansion* from the unimodal teacher.

timodal data is required). As illustrated in Figure 2, a unimodal network pre-trained on the labeled dataset plays the role of a teacher and distills information to a multimodal network, termed as a student. We observe an interesting phenomenon: our multimodal student, trained only on pseudo labels provided by the unimodal teacher, consistently outperforms the teacher under our training framework. We term this observation as knowledge expansion. Namely, a multimodal student is capable of refining pseudo labels. We conduct experimental results on various tasks and different modalities to verify this observation. We further offer empirical and theoretical explanations to understand the expansion capability of a multimodal student.

A closely related setting to ours is semi-supervised learning (SSL), whose goal is to improve a model's performance by leveraging unlabeled data of the same source, including modality. Different from SSL, we aim to develop an additional multimodal network on an unlabeled dataset. Despite the differences in modalities, *MKE* bears some similarity to SSL in terms of the mechanism. We provide a new perspective in addressing confirmation bias, a traditionally bothering problem in SSL. This bias stems from using incorrect predictions on unlabeled data for training and results in marginal performance gain over the original teacher network [3]. In SSL, various methods, *i.e.*, data augmentation [34, 43], injecting noise [44], meta-learning [29] have been proposed to address it. This work provides a novel angle orthogonal to these techniques in alleviating confirmation bias, by resorting to multimodal information. We demonstrate that multimodal inputs serve as a strong regularization, which helps correct inaccurate pseudo labels and overcome the limitation of unimodal networks.

## 2. Related Work

### 2.1. Semi-supervised Learning

**Pseudo labeling**, also known as self-training, is a simple and powerful technique in SSL, leading to great improvements on tasks such as image classification [23, 45, 44, 29], semantic segmentation [51, 10] and domain adaptation [52, 22]. One important limitation of pseudo labeling is confirmation bias [3]. Since pseudo labels are inaccurate, the student network may potentially learn these mistakes. Various works have been proposed to alleviate this bias [52, 3, 44, 29] while their discussion is limited to unimodality. **Consistency regularization** is another important brand of SSL. Based on model smoothness assumption, model predictions are constrained to be invariant to small perturbations of either inputs or model hidden states. A series of works have been proposed on producing random perturbations, such as using an exponential moving average of model parameters [36], data augmentation [43, 34], dropout [5, 44] or adversarial perturbations [27]. Recent works [7, 6, 34] combine consistency regularization with pseudo labeling together and demonstrate great benefits.

### 2.2. Cross-modal Distillation

Knowledge distillation (KD) [18] is an effective technique in transferring information from one network to another. KD has been broadly applied to model compression, where a lightweight student network learns from a cumbersome teacher network [40, 18, 46, 31, 37]. Another important application of KD is cross-modal distillation, where a teacher network transfers knowledge from one modality to a student learning from another modality. Various works have been proposed along this direction [15, 19, 4, 28, 2, 50, 49]. Gupta *et al*. [15] proposes a framework that transfers supervision from labeled RGB images to unlabeled depth and optical flow images. SoundNet [4] learns sound representations from well established visual recognition models using unlabeled videos. Zhao *et al*. [50] introduces an approach that estimates human poses using radio signals with cross-modal supervision signals provided by vision models.

### 2.3. Multimodal Learning

Models fusing data from multiple modalities has shown superior performance over unimodal models in various applications, for instance, sentiment analysis [48, 26], emotion recognition [38, 30], semantic segmentation [13, 39, 12, 41] and event classification [1]. One recent work [20] provides theoretical justifications on the advantages of multimodal learning over learning with single modality.

We compare our problem setting with prior works in Table 1. SSL adopts data from the same modality. Cross-modal distillation has the same training data assumption as us while they focus on testing with unimodal data only. Su-

| Related works | Train | | | | Test | |
|---|---|---|---|---|---|---|
| | labeled | | unlabeled | | | |
| | UM | MM | UM | MM | UM | MM |
| Semi-supervised learning | ✓ | | ✓ | | ✓ | |
| Cross-modal distillation | ✓ | | | ✓ | ✓ | |
| Supervised multimodal learning | | ✓ | | | | ✓ |
| MKE (ours) | ✓ | | | ✓ | | ✓ |

Table 1: Comparison of our data assumption with prior works. UM and MM denotes unimodality and multimodality respectively.

pervised multimodal learning does not take unlabeled data into consideration. Contrary to them, this work discusses a novel and practical scenario where only labeled unimodal and unlabeled multimodal data are available.

## 3. Approach

### 3.1. Multimodal Knowledge Expansion

**Problem formulation.** Without loss of generality, we limit our discussion to two modalities, denoted as $\alpha$ and $\beta$, respectively. We assume that a collection of labeled unimodal data $D_l = \{(\mathbf{x}_i^\alpha, \mathbf{y}_i)\}_{i=1}^N$ is given. Each sample input $\mathbf{x}_i^\alpha$ has been assigned a one-hot label vector $\mathbf{y}_i = \{0,1\}^K \in \mathcal{R}^K$, where $K$ is the number of classes. Besides the labeled dataset, an unlabeled multimodal dataset $D_u = \{(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta)\}_{i=1}^M$ is available. Our goal is to train a network parameterized by $\theta$ (*i.e.*, $\mathbf{f}(\mathbf{x};\theta)$) that could accurately predict the label $\mathbf{y}$ when its feature $\mathbf{x} = (\mathbf{x}^\alpha, \mathbf{x}^\beta)$ is given.

To transfer the knowledge of a labeled unimodal dataset $D_l$ to an unlabeled multimodal dataset $D_u$, we present a simple and efficient model-agnostic framework named multimodal knowledge expansion (*MKE*) in Algorithm 1. We first train a unimodal teacher network $\theta_t^\star$ on the labeled dataset $D_l$. Next, the obtained teacher is employed to generate pseudo labels for the multimodal dataset $D_u$, yielding $\tilde{D}_u$. Finally, we train a multimodal student $\theta_s^\star$ based on the pseudo-labeled $\tilde{D}_u$ with the loss term described in Equation (3)-(5).

In order to prevent the student from confirming to teacher's predictions (*i.e.*, confirmation bias [3]), the loss term in Equation (3)-(5) has been carefully designed. It combines the standard pseudo label loss (*i.e.*, Equation (4)) and a regularization loss (*i.e.*, Equation (5)). Intuitively speaking, pseudo label loss aims to minimize the difference between a multimodal student and the unimodal teacher, while regularization loss enforces the student to be invariant to small perturbations of input or hidden states. In the context of multimodal learning, the regularization term encourages the multimodal student to learn from the information brought by the extra modality $\beta$, and meanwhile, ensures that the student does not overfit to teacher's predictions based solely on modality $\alpha$. Note that in our implementation, to avoid introducing and tuning one extra hyperparameter $\gamma$ and save computation time, we train the student network with $\theta_s^\star = \text{argmin}_{\theta_s} \frac{1}{M}\sum_{i=1}^M l_{cls}(\tilde{\mathbf{y}}_i, \mathcal{T}(\mathbf{f}_s(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta;\theta_s))$, which is equivalent to Equation (3). The detailed proof is provided in the supplementary material.

**An illustrative example.** We consider a variant of the 2D-TwoMoon [3] problem shown in Figure 3a. The data located at the upper moon and lower moon have true labels 0 and 1, and are colored by red and blue, respectively. The deeply blue- or red-colored large dots compose the labeled unimodal dataset $D_l$, and only their X coordinates are known. On the other hand, $D_u$ consists of all lightly-colored small dots, with both X and Y coordinates available. Namely, modality $\alpha$ and $\beta$ are interpreted as observing from the X-axis and Y-axis, respectively.

**Algorithm 1** multimodal knowledge expansion (*MKE*)

(1) Train a unimodal teacher $\theta_t^\star$ with the labeled dataset $D_l = \{(\mathbf{x}_i^\alpha, \mathbf{y}_i)\}_{i=1}^N$:

$$\theta_t^\star = \underset{\theta_t}{\text{argmin}} \frac{1}{N}\sum_{i=1}^N l_{cls}(\mathbf{y}_i, \mathbf{f}_t(\mathbf{x}_i^\alpha;\theta_t)) \quad (1)$$

(2) Generate pseudo labels for $D_u = \{(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta)\}_{i=1}^M$ by using the teacher model $\theta_t^\star$, yielding the pseudo-labeled dataset $\tilde{D}_u = \{(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta, \tilde{\mathbf{y}}_i)\}_{i=1}^M$:

$$\tilde{\mathbf{y}}_i = \mathbf{f}_t(\mathbf{x}_i^\alpha;\theta_t^\star), \forall\, (\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta) \in D_u \quad (2)$$

(3) Train a multimodal student $\theta_s^\star$ with $\tilde{D}_u$:

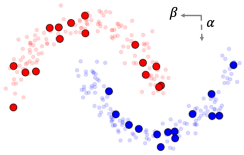
$$\theta_s^\star = \underset{\theta_s}{\text{argmin}}(\mathcal{L}_{pl} + \gamma\mathcal{L}_{reg}) \quad (3)$$

$$\mathcal{L}_{pl} = \frac{1}{M}\sum_{i=1}^M l_{cls}(\tilde{\mathbf{y}}_i, \mathbf{f}_s(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta;\theta_s)) \quad (4)$$

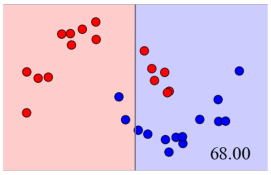
$$\mathcal{L}_{reg} = \sum_{i=1}^M l_{reg}[\mathbf{f}_s(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta;\theta_s), \mathcal{T}(\mathbf{f}_s(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta;\theta_s))] \quad (5)$$

$l_{cls}$: cross entropy loss for hard $\tilde{\mathbf{y}}_i$ and KL divergence loss for soft $\tilde{\mathbf{y}}_i$.
$l_{reg}$: a distance metric (*e.g.*, L2 norm).
$\gamma$: a constant balancing the weight of $\mathcal{L}_{pl}$ and $\mathcal{L}_{reg}$.
$\mathcal{T}$: a transformation defined on the student model, realized via input or model perturbations (*i.e.*, augmentations, dropout).

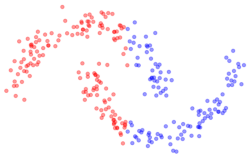
We first train a teacher with the labeled unimodal dataset $D_l$. The learned classification boundary is demonstrated in Figure 3b. Next, we adopt the learned teacher to generate pseudo labels for $D_u$. As indicated in Figure 3c, pseudo labels may be inaccurate and disagree with ground truth: in our toy example, the unimodal teacher only yields 68% accuracy. As shown in Figure 3f, provided with these not-so-accurate pseudo labels, the student could still outperform the teacher by a large margin (*i.e.*, about 13% more accurate). It presents a key finding in our work: *Despite no*



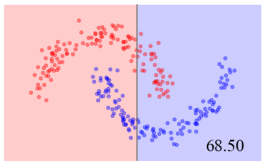

(a) Deeply-colored large dots observed from the X-axis compose $D_l$ and lightly-colored small dots observed from the XY-plane compose $D_u$.



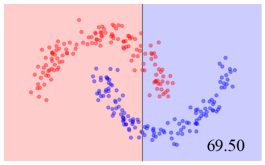

(b) Train a unimodal classifier on labeled data points. Since only X coordinates of the data in $D_l$ are known, it is natural that the boundary is vertical.



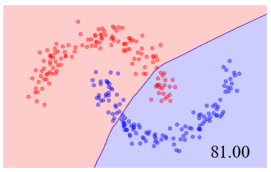
(c) The teacher network generates pseudo labels for unlabeled data




(d) Naive pseudo labeling overfits to inaccurate pseudo labels




(e) Consistency regularization slightly improves a unimodal student




(f) *MKE* greatly improves a multimodal student

Figure 3: (a)-(c) problem description and illustration of *MKE* using the TwoMoon example; (d)-(f) comparison of naive pseudo labeling, consistency training methods, and the proposed *MKE*. Values in the bottom right corner denotes test accuracy (%).

*access to ground truth, a multimodal student is capable of correcting inaccurate labels and outperforms the teacher network. Knowledge expansion is achieved.*

### 3.2. Rectifying Pseudo Labels

The somewhat surprising finding about knowledge expansion further motivates our thinking: *where does the expansion capability of a multimodal student come from?* In this section, we will answer this question with the TwoMoon example.

To start with, we consider directly adopting unimodal SSL for this problem. Namely, given a teacher network $\theta_t^\star$ trained with labeled data $D_l$ and an unlabeled multi-modal dataset $D_u$, the student network takes $\mathbf{x}_i^\alpha \in D_u$ as input. Naive pseudo labeling [23] uses the following loss to minimize the disagreement between the fixed teacher $\theta_t^\star$ and a student network $\theta_s$:

$$\mathcal{L}'_{pl} = \mathbb{E}_{\mathbf{x}_i^\alpha \in D_u}\{l_{cls}[\mathbf{f}_t(\mathbf{x}_i^\alpha;\theta_t^\star), \mathbf{f}_s(\mathbf{x}_i^\alpha;\theta_s)]\} \quad (6)$$

However, due to confirmation bias [3], the student network is likely to overfit to incorrect pseudo labels provided by the teacher network, yielding $\mathbf{f}_s(\mathbf{x};\theta_s^\star)$ similar to $\mathbf{f}_t(\mathbf{x};\theta_t^\star)$, if not identical. In the TwoMoon example, we observe that the unimodal student trained with Equation (6) achieves similar performance as its teacher. This is demonstrated in Figure 3d.

To address this bias, we follow the thought of consistency training methods in SSL [27, 43, 34] and introduce one general regularization loss term to enforce model smoothness:

$$\mathcal{L}'_{reg} = \mathbb{E}_{\mathbf{x}_i^\alpha \in D_u}\{l_{reg}[\mathbf{f}_s(\mathbf{x}_i^\alpha;\theta_s), \mathcal{T}'(\mathbf{f}_s(\mathbf{x}_i^\alpha;\theta_s))]\} \quad (7)$$

Namely, $\mathcal{L}'_{reg}$ encourages the model to output similar predictions for small perturbations of the input or the model. $\mathcal{T}'(\mathbf{f}_s(\mathbf{x}_i^\alpha;\theta_s))$ denotes transformation applied to unimodal inputs or model hidden states, which can be realized via input augmentation, noise, dropout, etc. As shown in Figure 3e, the unimodal student trained with a combined loss of Equation (6)-(7) achieves about 69.50% prediction accuracy. While it indeed outperforms the teacher of 68.00% accuracy shown in Figure 3b, the unimodal student under consistency regularization fails to utilize unlabeled data effectively and only brings marginal improvement. Although confirmation bias is slightly reduced by the regularization term in Equation (7), it still heavily constrains performance of unimodal SSL methods.

Therefore, we turn to multimodality as a solution and resort to the information brought by modality $\beta$. Utilizing both modalities in $D_u$, we substitute unimodal inputs shown in Equation (6)-(7) with multimodal ones and derive the loss terms for training a multimodal student:

$$\mathcal{L}_{pl} = \mathbb{E}\{l_{cls}[\mathbf{f}_t(\mathbf{x}_i^\alpha;\theta_t^\star), \mathbf{f}_s(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta;\theta_s)]\} \quad (8)$$

$$\mathcal{L}_{reg} = \mathbb{E}\{l_{reg}[\mathbf{f}_s(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta;\theta_s), \mathcal{T}(\mathbf{f}_s(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta;\theta_s))]\} \quad (9)$$

where both expectations are performed with respect to $(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta) \in D_u$. In fact, Equation (8)-(9) reduces to Equation (4)-(5) when $D_u$ is a finite set containing $M$ multimodal samples. As shown in Figure 3f, we observe substantial improvement of a multimodal student (*i.e.*, 81.00% accuracy) over the teacher (*i.e.*, 68.00% accuracy). It implies that a multimodal student effectively alleviates confirmation bias and leads to superior performance over the teacher.

To understand the principles behind this phenomenon, we train one unimodal student with Equation (6)-(7) and one multimodal student with Equation (8)-(9) on the TwoMoon data. Transformation $\mathcal{T}$ is defined on model inputs and implemented as additive Gaussian noise. Figure 4 visualizes the transformation space of one data sample A with both pseudo label and true label being "red". Data B is one point that the teacher predicts "blue" while its true label is "red". The pseudo label and true label of data C are "blue".

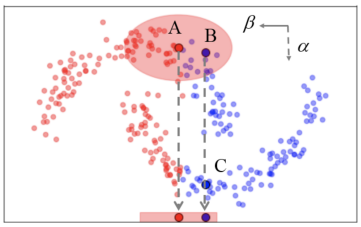



Figure 4: Illustration of the transformation space of one data sample A. The 1-D red line on X-axis corresponds to the transformation space of a unimodal student while the 2-D red circle corresponds to that of a multimodal student.

When training a unimodal student, we only know the X coordinates of data points, and the transformation space defined by $\mathcal{T}'$ is given by the 1-D red line on X-axis. Under this circumstance, minimizing $\mathcal{L}'_{reg}$ in Equation (7) encourages the unimodal student to predict label "red" for the data point located in the red line. This is the case for B, but it will also flip the teacher's prediction for C and make it wrong! The intrinsic reason is that restricted by unimodal inputs, the student network can not distinguish along the Y-axis and mistakenly assumes that C locates near A.

On the contrary, the extra modality $\beta$ helps us see the real distances among A, B, and C. Transformation space of data A in the case of a multimodal student is given by the red circle in Figure 4. A multimodal student is guided to predict "red" for data falling inside the circle. This time B locates in the transformation space, while C doesn't. Therefore, the multimodal student can correct the wrong pseudo label of data B due to the regularization constraint in Equation (9), and its decision boundary is pushed closer to the ground truth. This example demonstrates that multimodality serves as a strong regularization and enables the student to "see" something beyond the scope of its teacher, resulting in knowledge expansion.

### 3.3. Theoretical Analysis

In this section, we provide a theoretical analysis of *MKE*. Building upon unimodal self-training [42], we prove that our multimodal student improves over pseudo labels given by the teacher.

Consider a $K$-way classification problem, and assume that we have a teacher network pre-trained on a collection of labeled data $D_l$. We further assume a set of unlabeled multimodal data $D_u = \{\mathbf{x}_i = (\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta) \in \mathcal{X}\}_{i=1}^M$ is available, where $\mathcal{X} = \mathcal{X}^\alpha \times \mathcal{X}^\beta$. Let $\mathbf{f}_\star(\mathbf{x};\theta_\star), \mathbf{f}_t(\mathbf{x};\theta_t), \mathbf{f}_s(\mathbf{x};\theta_s)$ denote the ground truth classifier, a teacher classifier, and a student classifier, respectively. Error of an arbitrary classifier $\mathbf{f}(\mathbf{x};\theta)$ is defined as: $Err(\mathbf{f}(\mathbf{x};\theta)) = \mathbb{E}_{\mathbf{x}}[\mathbf{f}(\mathbf{x};\theta) \neq \mathbf{f}_\star(\mathbf{x};\theta_\star)]$. Let $P$ refer to a distribution of unlabeled samples over input space $\mathcal{X}$. $P_i$ denotes the class-conditional distribution of $\mathbf{x}$ conditioned on $\mathbf{f}_\star(\mathbf{x};\theta_\star) = i$. We use $\mathcal{M}(\theta_t) \subseteq D_u$ to denote the set of multimodal data that the teacher gives wrong predictions on, *i.e.*, $\mathcal{M}(\theta_t) = \{(\mathbf{x}^\alpha, \mathbf{x}^\beta) | \mathbf{f}_t(\mathbf{x}^\alpha;\theta_t) \neq \mathbf{f}_\star(\mathbf{x}^\alpha;\theta_\star), (\mathbf{x}^\alpha, \mathbf{x}^\beta) \in D_u\}$. Let $\bar{a} = max_i\{P_i(\mathcal{M}(\theta_t)\}$ refer to the maximum fraction of data misclassified by the teacher network in any class.

We first require data distribution $P$ to satisfy the following expansion assumption, which states that data distribution has good continuity in input spaces.

**Assumption 1** *$P$ satisfies $(\bar{a}, c_1)$ and $(\bar{a}, c_2)$ expansion [42] on $\mathcal{X}^\alpha$ and $\mathcal{X}^\beta$, respectively, with $1 < min(c_1, c_2) \leq max(c_1, c_2) \leq \frac{1}{\bar{a}}$ and $c_1 c_2 > 5$.*

$$P_i(N(V^\alpha)) \geq min\{c_1 P_i(V^\alpha), 1\}, \quad \forall\, i \in [K], \forall\, V^\alpha \subseteq \mathcal{X}^\alpha \text{ with } P_i(V^\alpha) \leq \bar{a} \tag{10}$$

$$P_i(N(V^\beta)) \geq min\{c_2 P_i(V^\beta), 1\}, \quad \forall\, i \in [K], \forall\, V^\beta \subseteq \mathcal{X}^\beta \text{ with } P_i(V^\beta) \leq \bar{a} \tag{11}$$

where $N(V)$ denotes the neighborhood of a set $V$, following the same definition as in [42].

Furthermore, we assume conditional independence of multimodal data in Assumption 2, which is widely adopted in the literature of multimodal learning [24, 8, 35].

**Assumption 2** *Conditioning on ground truth labels, $\mathcal{X}^\alpha$ and $\mathcal{X}^\beta$ are independent.*

$$P_i(V^\alpha, V^\beta) = P_i(V^\alpha) \cdot P_i(V^\beta), \quad \forall\, i \in [K], \forall\, V^\alpha \subseteq \mathcal{X}^\alpha, \forall\, V^\beta \subseteq \mathcal{X}^\beta \tag{12}$$

**Lemma 1** *Data distribution $P$ on $\mathcal{X}$ satisfies $(\bar{a}, c_1 c_2)$ expansion.*

Proof of Lemma 1 is provided in the supplementary material. We state below that the error of a multimodal student classifier is upper-bounded by the error of its teacher. We follow the proof in [42] to prove Theorem 1.

**Theorem 1** *Suppose Assumption 3.3 of [42] holds, a student classifier $\mathbf{f}_s(\mathbf{x}^\alpha, \mathbf{x}^\beta;\theta_s)$ that minimizes loss in Equation (3) (in the form of Equation 4.1 of [42]) satisfies:*

$$Err(\mathbf{f}_s(\mathbf{x}^\alpha, \mathbf{x}^\beta;\theta_s)) \leq \frac{4 \cdot Err(\mathbf{f}_t(\mathbf{x}^\alpha;\theta_t))}{c_1 c_2 - 1} + 4\mu \tag{13}$$

where $\mu$ appears in Assumption 3.3 of [42] and is expected to be small or negligible. Theorem 1 helps explain the empirical finding about knowledge expansion. Training a multimodal student $\mathbf{f}(\mathbf{x}^\alpha, \mathbf{x}^\beta;\theta_s)$ on pseudo labels given by a pre-trained teacher network $\mathbf{f}(\mathbf{x}^\alpha;\theta_t)$ refines pseudo labels.

In addition, the error bound of a unimodal student $\mathbf{f}_s(\mathbf{x}^\alpha;\theta_s)$ that only takes inputs from modality $\alpha$ and pseudo labels is given by:

$$Err(\mathbf{f}_s(\mathbf{x}^\alpha;\theta_s)) \leq \frac{4 \cdot Err(\mathbf{f}_t(\mathbf{x}^\alpha;\theta_t))}{c_1 - 1} + 4\mu \tag{14}$$

By comparing Equation (13) and (14), we observe that the role of multimodality is to increase the expansion factor from $c_1$ to $c_1 c_2$ and to improve the accuracy bound. This observation further confirms our empirical finding and unveils the role of *MKE* in refining pseudo labels from a theoretical perspective.

# 4. Experimental Results

To verify the efficiency and generalizability of the proposed method, we perform a thorough test of *MKE* on various tasks: (i) binary classification on the synthetic TwoMoon dataset, (ii) emotion recognition on RAVDESS [25] dataset, (iii) semantic segmentation on NYU Depth V2 [32] dataset, and (iv) event classification on AudioSet [14] and VGGsound [9] dataset. We emphasize that the above four tasks cover a broad combination of modalities. For instance, modalities $\alpha$ and $\beta$ represent images and audios in (ii), where images are considered as a "weak" modality in classifying emotions than images. In (iii), modality $\alpha$ and $\beta$ refer to RGB and depth images, respectively, where RGB images play a central role in semantic segmentation and depth images provide useful cues.

**Baselines.** Our multimodal student (termed as MM student) trained with *MKE* is compared with the following baselines:

- UM teacher: a unimodal teacher network trained on $(\mathbf{x}^\alpha, \mathbf{y}_i) \in D_l$.
- UM student: a unimodal student network trained on $(\mathbf{x}^\alpha, \tilde{\mathbf{y}}_i) \in \tilde{D}_u$ (*i.e.*, uni-modal inputs and pseudo labels given by the UM teacher).
- NOISY student [44]: a unimodal student network trained on $(\mathbf{x}^\alpha, \mathbf{y}_i) \in D_l \cup (\mathbf{x}^\alpha, \tilde{\mathbf{y}}_i) \in \tilde{D}_u$ with noise injected during training.
- MM student (no reg): a multimodal student network trained with no regularization (*i.e.*, Equation (5) is not applied during training).
- MM student (sup): a multimodal student trained on $D_u$ with true labels provided. This supervised version can be regarded as the upper bound of our multimodal student.

Since iterative training [44] can be applied to other baselines and our MM student as well, the number of iterations of a NOISY student is set as one to ensure a fair comparison. We employ different regularization techniques as $\mathcal{T}$ in Equation (5) for the four tasks to demonstrate the generalizability of our proposed methods. Regularization is applied to all baselines identically except for MM student (no reg).

Furthermore, we present an ablation study of various components of *MKE*, *i.e.*, unlabeled data size, teacher model, hard *vs.* soft labels, along with dataset and implementation details in the supplementary material.

## 4.1. TwoMoon Experiment

We first provide results on synthetic TwoMoon data. We generate 500 samples making two interleaving half circles, each circle corresponding to one class. The dataset is randomly split as 30 labeled samples, 270 unlabeled samples and 200 test samples. X and Y coordinates of data are interpreted as modality $\alpha$ and $\beta$, respectively.

**Baselines & Implementation.** We implement both the UM teacher and the UM student networks as 3-layer MLPs with 32 hidden units, while the MM student has 16 hidden units. We design three kinds of transformations $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$ used in Equation (5): (i) $\mathcal{T}_1$: adding zero-mean Gaussian noise to the input with variance $v_0$, (ii) $\mathcal{T}_2$: adding zero-mean Gaussian noise to outputs of the first hidden layer with variance $v_1$, and (iii) $\mathcal{T}_3$: adding a dropout layer with dropout rate equal to $r_0$. By adjusting the values of $v_0$, $v_1$ and $r_0$, we could test all methods under no / weak / strong regularization. Specifically, higher values indicate stronger regularization.

| Methods | Test Accuracy (%) | | |
|---|---|---|---|
| UM teacher | 68.00 | | |
| $\mathcal{T}_1$ | $v_0 = 0$ | $v_0 = 1$ | $v_0 = 2$ |
| UM student | 68.00 | 69.90 | 72.80 |
| MM student (ours) | 68.85 | 80.75 | **83.15** |
| MM student (sup) | 88.05 | 87.35 | 86.95 |
| $\mathcal{T}_2$ | $v_1 = 0$ | $v_1 = 5$ | $v_1 = 10$ |
| UM student | 68.00 | 68.95 | 70.05 |
| MM student (ours) | 68.85 | 80.00 | **82.10** |
| MM student (sup) | 88.05 | 87.40 | 86.40 |
| $\mathcal{T}_3$ | $r_0 = 0$ | $r_0 = 0.4$ | $r_0 = 0.8$ |
| UM student | 68.00 | 68.40 | 68.95 |
| MM student (ours) | 68.85 | 73.65 | **79.20** |
| MM student (sup) | 88.05 | 87.35 | 86.90 |

Table 2: Results of TwoMoon experiment. A MM student significantly outperforms a UM student and teacher under consistency regularization.

**Results.** Table 2 demonstrates that a MM student under consistency regularization outperforms its unimodal counterpart in all cases of $\mathcal{T}$. Specifically, a MM student under strong regularization achieves closes results with MM student (sup), as shown in the last column. The small gap between a MM student (trained on pseudo labels) and its upper bound (trained on true labels) indicates the great expansion capability of *MKE*. In addition, we observe better performance of both UM and MM student with increasing regularization strength, demonstrating that consistency regularization is essential in alleviating confirmation bias.

### 4.2. Emotion Recognition

We evaluate *MKE* on RAVDESS [25] dataset for emotion recognition. The dataset is randomly split as 2:8 for $D_l$ and $D_u$ and 8:1:1 as train / validation / test for $D_u$. Images and audios are considered as modality $\alpha$ and $\beta$, respectively.

**Baselines & Implementation.** For the MM student, we adopt two 3-layer CNNs to extract image and audio features, respectively. The two visual and audio features are concatenated into a vector and then passed through a 3-layer MLP. The UM teacher, UM student and NOISY student are identical to the image branch of a MM student network, also followed by a 3-layer MLP. $\mathcal{T}$ in Equation (5) is implemented as one dropout layer of rate 0.5.

**Results.** As shown in Table 3, with the assistance of labeled data and consistency regularization, NOISY student generalizes better than the UM teacher and UM student, achieving 83.09% accuracy over 80.33% and 77.79%. Still, the improvement is trivial. In contrast, our MM student network improves substantially over the original teacher network despite no access to ground truth and leads to 91.38% test accuracy. The great performance gain can be attributed to additional information brought by audio modality. It demonstrates that *MKE* can be plugged into existing SSL methods like NOISY student for boosting performance when multimodal data are available. Furthermore, regularization helps our MM student yield better performance than the MM student (no reg). More results are presented in the supplementary material.

| Methods | Train data | | | Accuracy (%) | |
|---|---|---|---|---|---|
| | *mod* | $D_l$ | $\tilde{D}_u$ | val | test |
| UM teacher | $i$ | ✓ | | 79.67 | 80.33 |
| UM student | $i$ | | ✓ | 79.01 | 77.79 |
| NOISY student [44] | $i$ | ✓ | ✓ | 82.54 | 83.09 |
| MM student (no reg) | $i, a$ | | ✓ | 88.73 | 89.28 |
| MM student (ours) | $i, a$ | | ✓ | 90.61 | **91.38** |
| MM student (sup) | $i, a$ | | $\star$ | 97.46 | 97.35 |

Table 3: Results of emotion recognition on RAVDESS. *mod*, $i$ and $a$ denote modality, images and audios, respectively. Data used for training each method is listed. $\star$ means that the MM student (sup) is trained on true labels instead of pseudo labels in $\tilde{D}_u$.

### 4.3. Semantic Segmentation

We evaluate our method on NYU Depth V2 [32]. It contains 1449 RGB-D images with 40-class labels, where 795 RGB images are adopted as $D_l$ for training the UM teacher and the rest 654 RGB-D images are for testing. Besides labeled data, NYU Depth V2 also provides unannotated video sequences, where we randomly extract 1.5K frames of RGB-D images as $D_u$ for training the student. Modality $\alpha$ and $\beta$ represents RGB images and depth images.

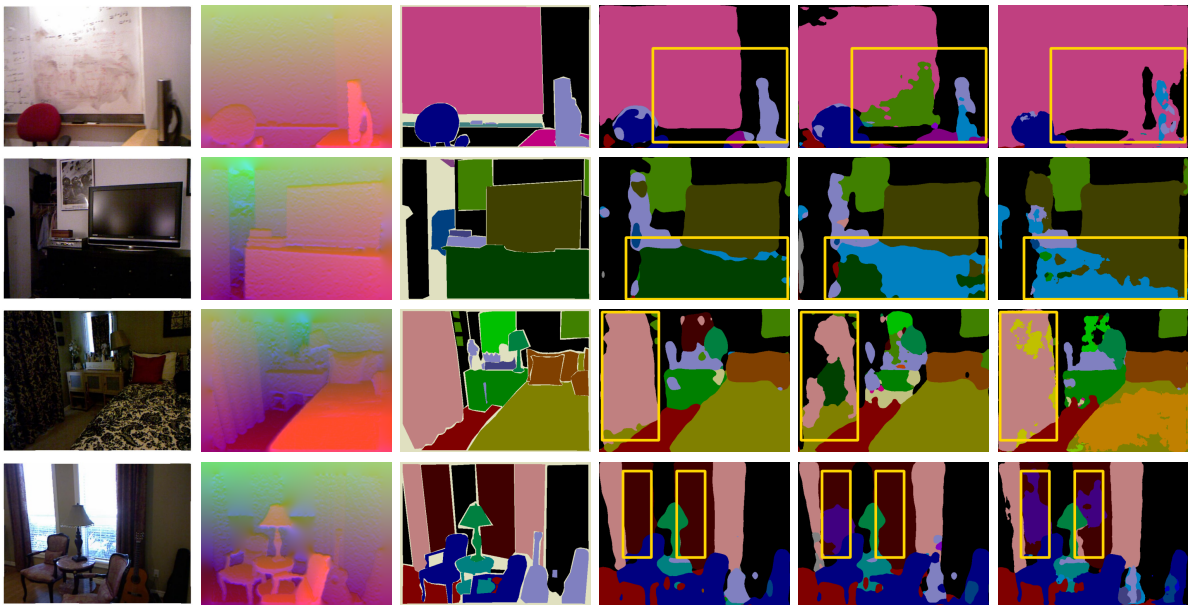
| Method | Train data | | | Test mIoU |
|---|---|---|---|---|
| | *mod* | $D_l$ | $\tilde{D}_u$ | (%) |
| UM teacher | $rgb$ | ✓ | | 44.15 |
| Naive student [10] | $rgb$ | | ✓ | 46.13 |
| NOISY student [44] | $rgb$ | ✓ | ✓ | 47.68 |
| Gupta *et al.* [15] | $rgb, d$ | | ✓ | 45.65 |
| CMKD [49] | $rgb, d$ | | ✓ | 45.25 |
| MM student (no reg) | $rgb, d$ | | ✓ | 46.14 |
| MM student (ours) | $rgb, d$ | | ✓ | **48.88** |

Table 4: Results of semantic segmentation on NYU Depth V2. $rgb$ and $d$ denote RGB images and depth images.

**Baselines & Implementation.** We compare *MKE* against SSL methods [10] [44] and cross-modal distillation [15] [49]. Due to different problem settings, we slightly modified cross-modal distillation methods to make them comparable. Since RGB-D images from $D_u$ are unannotated, we are unable to train a supervised version of the MM student (*i.e.*, MM student (sup)) in this task. We adopt ResNet-101 [16] as backbone and DeepLab V3+ [11] as decoder for the UM teacher. In terms of training a MM student, depth images are first converted to HHA images and then passed to a fusion network architecture proposed in [12] along with RGB images. We design the UM student architecture as the RGB branch of a MM student network. For the regularization term, we employ input augmentation for RGB images, *i.e.*, random horizontal flipping and scaling with scales [0.5,1.75].

**Results.** Table 4 reports mean Intersection-over-Union (mIoU) of each method. We observe that a MM student greatly improves over the UM teacher, *i.e.*, achieves a mIoU of 48.88 % while it is trained on pseudo labels of approximately 44.15% mIoU. Furthermore, provided with no ground truth, our MM student outperforms all baselines with a considerable performance gain. This demonstrates the effectiveness of *MKE*. We also arrive at the same conclusion that regularization helps improve the MM student since our MM student yields higher accuracy than a MM student (no reg). It indicates that *MKE* and current SSL methods that focus on designing augmentations to emphasize consistency regularization can be combined together to boost performance.

Visualization results in Figure 5 demonstrate that our MM student refines pseudo labels and achieves knowledge expansion. Although it receives noisy predictions given by the UM teacher, our MM student does a good job in handling details and maintaining intra-class consistency. As shown



(a) RGB images (b) depth images (c) ground truth (d) MM student (ours) (e) UM teacher (f) NOISY student

Figure 5: Qualitative segmentation results on NYU Depth V2 test set.

in the third and fourth row, the MM student is robust to illumination changes while the UM teacher and NOISY student easily get confused. Depth modality helps our MM student better distinguish objects and correct wrong predictions it receives. More qualitative examples are shown in the supplementary material.

### 4.4. Event Classification

We present experimental results on a real-world application, event classification. 3.7K audios from AudioSet [14] and 3.7K audio-video pairs from VGGSound [9] are taken as the labeled unimodal dataset $D_l$ and unlabeled multimodal dataset $D_u$, respectively. In this task, modality $\alpha$ and $\beta$ correspond to audios and videos.

**Baslines & Implementation.** For the UM teacher, we take ResNet-18 as the backbone and a linear layer as classification layer. For the MM student, the audio backbone is identical to that of the UM teacher, and the video backbone is a ResNet-18 with 3D convolution layers. Features from the audio and video backbone are concatenated together before feeding into one classification layer. Following the same regularization term of [9], we randomly sample audio clips of 5 seconds and apply short-time Fourier Transformation for $257 \times 500$ spectrograms during training.

**Results.** Table 5 reports mean Average Precision (mAP) of each method. The baseline model is the UM teacher trained on $D_l$, which achieves a 0.345 mAP. Benefiting from the video modality, our MM student achieves best performance with a mAP of 0.427, outperforming NOSIY student [44] and cross-modal distillation methods [28] [49]. Notably, the difference between our MM student and its upper bound (*i.e.*, MM student (sup)) is small, showing great potentials of *MKE* in correcting pseudo labels.

| Method | Train data | | | Test mAP |
|---|---|---|---|---|
| | *mod* | $D_l$ | $\tilde{D}_u$ | |
| UM teacher | $a$ | ✓ | | 0.345 |
| UM student | $a$ | | ✓ | 0.406 |
| NOISY student [44] | $a$ | ✓ | ✓ | 0.411 |
| Owens *et al.* [28] | $a, v$ | | ✓ | 0.371 |
| CMKD [49] | $a, v$ | | ✓ | 0.372 |
| MM student (no reg) | $a, v$ | | ✓ | 0.421 |
| MM student (ours) | $a, v$ | | ✓ | **0.427** |
| MM student (sup) | $a, v$ | | ⋆ | 0.434 |

Table 5: Results of event classification on AudioSet and VGGSound. $a$ and $v$ indicate audios and videos.

## 5. Conclusion

Motivated by recent progress on multimodal data collection, we propose a multimodal knowledge expansion framework to effectively utilize abundant unlabeled multimodal data. We provide theoretical analysis and conduct extensive experiments, demonstrating that a multimodal student corrects inaccurate predictions and achieves knowledge expansion from the unimodal teacher. In addition, compared with current semi-supervised learning methods, *MKE* offers a novel angle in addressing confirmation bias.

## References


[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *CVPR*, pages 14679–14689, 2020. 2

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, pages 609–617, 2017. 2

[3] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IJCNN*, pages 1–8. IEEE, 2020. 2, 3, 4

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *NeurIPS*, 2016. 2

[5] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pages 3365–3373, 2014. 2

[6] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *ICLR*, 2020. 2

[7] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 2

[8] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998. 5

[9] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725. IEEE, 2020. 6, 8

[10] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *ECCV*, pages 695–714, 2020. 2, 7

[11] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 7

[12] Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. *ECCV*, 2020. 2, 7

[13] Di Feng, Christian Haase-Schuetz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 2

[14] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, pages 776–780. IEEE, 2017. 6, 8

[15] Saurabh Gupta et al. Cross modal distillation for supervision transfer. In *CVPR*, 2016. 2, 7

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 7

[17] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, pages 131–135. IEEE, 2017. 1

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[19] Judy Hoffman, Saurabh Gupta, Jian Leong, Sergio Guadarrama, and Trevor Darrell. Cross-modal adaptation for rgb-d detection. In *ICRA*, pages 5032–5039. IEEE, 2016. 2

[20] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multimodal learning better than single (provably). *arXiv preprint arXiv:2106.04538*, 2021. 2

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25:1097–1105, 2012. 1

[22] Ananya Kumar et al. Understanding self-training for gradual domain adaptation. In *ICML*, 2020. 2

[23] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013. 2, 4

[24] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *ECCV*, pages 4–15. Springer, 1998. 5

[25] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 6, 7

[26] Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems*, 161:124–133, 2018. 2

[27] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE TPAMI*, 41(8):1979–1993, 2018. 2, 4

[28] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, pages 801–816, 2016. 2, 8

[29] Hieu Pham, Qizhe Xie, Zihang Dai, and Quoc V Le. Meta pseudo labels. *arXiv preprint arXiv:2003.10580*, 2020. 2

[30] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. Multimodal emotion recognition using deep learning architectures. pages 1–9. IEEE, 2016. 2

[31] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2

[32] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012. 6, 7

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[34] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 2, 4

[35] Xinwei Sun, Yilun Xu, Peng Cao, Yuqing Kong, Lingjing Hu, Shanghang Zhang, and Yizhou Wang. Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In *ECCV*, pages 171–188, 2020. 5

[36] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 2

[37] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, pages 1365–1374, 2019. 2

[38] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017. 2

[39] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *IJCV*, pages 1–47, 2019. 2

[40] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE TPAMI*, 2021. 2

[41] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *NeurIPS*, 2020. 2

[42] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020. 5

[43] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 33, 2020. 2, 4

[44] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020. 2, 6, 7, 8

[45] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 2

[46] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan L Yuille. Training deep neural networks in generations: A more tolerant teacher educates better students. In *AAAI*, volume 33, pages 5628–5635, 2019. 2

[47] Dong Yu and Li Deng. *AUTOMATIC SPEECH RECOGNITION*. Springer, 2016. 1

[48] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *EMNLP*, 2017. 2

[49] Long Zhao et al. Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In *CVPR*, 2020. 2, 7, 8

[50] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *CVPR*, pages 7356–7365, 2018. 2

[51] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *NeurIPS*, 33, 2020. 2

[52] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, pages 5982–5991, 2019. 2