Multimodal Knowledge Expansion Supplementary Materials

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

[1]Shanghai Qi Zhi Institute, [2]The University of Texas at Austin, [3]South China University of Technology
[4]Massachusetts Institute of Technology, [5]Tsinghua University

This supplementary material presents: (1) dataset and implementation details; (2) more qualitative experimental results; (3) ablation studies; (4) proofs in Section 3.

1. Dataset and Implementation Details

1.1. Emotion Recognition

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains videos and audios of 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements. It contains 1440 emotional utterances with 8 different emotion classes: neutral, calm, happy, sad, angry, fearful, disgust and surprise. 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. To construct the labeled uni-modal dataset D_l, we select images every 0.5 second of a video clip as modality α and train a facial emotion recognition (FER) network as the UM teacher, which classifies emotions based on images. Image-audio pairs from video clips consist of the unlabeled multimodal dataset D_u. We sample images as inputs from modality α in the same way, adopt "Kaiser best" sampling for audios and take Mel-frequency cepstral coefficients (MFCCs) as inputs from modality β.

1.2. Semantic Segmentation

NYU Depth V2 contains 1449 RGB-D images with 40-class labels, where 795 RGB images are adopted 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. We randomly sample 1488 RGB-D images as D_u for training the student. Soft labels of the UM teacher are adopted.

In addition, we propose a confidence-weighted loss term in this task to further regularize the student, preventing it from overfiting to the teacher. For each sample pixel $\mathbf{x}$ and its soft pseudo label $\tilde{\mathbf{y}}$, we assign $\mathbf{x}$ with a weight $\omega(\mathbf{x})$ defined by:

$$\omega(\mathbf{x}) = 1 - \frac{\sum_{k=1}^{K} \tilde{\mathbf{y}}_i \log \tilde{\mathbf{y}}_i}{\log K} \tag{1}$$

K denotes the number of classes. We then modify $\mathcal{L}_{pl}$ in Equation (4) of the main paper by applying a weight for each sample:

$$\mathcal{L}_{pl} = \frac{1}{M} \sum_{i=1}^{M} \omega(\mathbf{x})\, l_{cls}(\tilde{\mathbf{y}}_i, \mathbf{f}_s(\mathbf{x}_i^{\alpha}, \mathbf{x}_i^{\beta}; \theta_s)) \tag{2}$$

Figure 2 demonstrates one image and its confidence map (*i.e.*, $\omega(\mathbf{x})$) based on pseudo labels of the UM teacher. Low confidence pixels are given a small weight while high confidence ones contribute largely in calculating the loss. This technique helps further reduce noise brought by inaccurate pseudo labels.

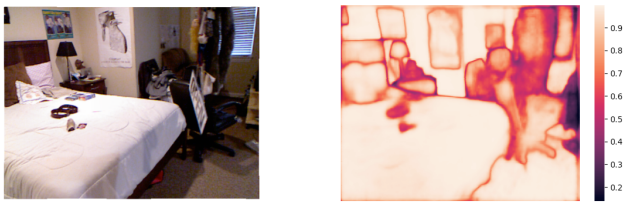
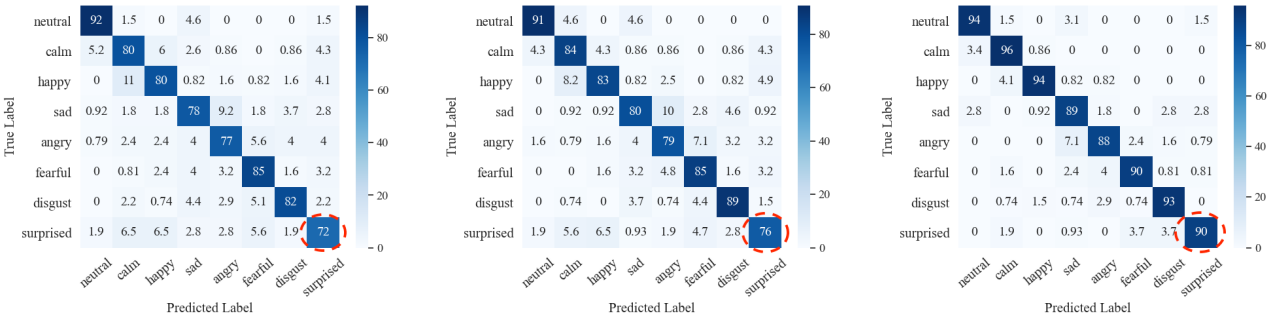


Figure 2: Left figure: one RGB image; right figure: corresponding weight value ω of each pixel.

1.3. Event Classification

The AudioSet and VGGSound are both audio-visual datasets for event classification. We take a mini common set of them including 3710 data in AudioSet and 3748 data for training and 1937 data for testing in VGGSound with 46 event categories. VGGSound guarantees the audio-video correspondence as the sound source is visually evident within the video, while AudioSet does not. Therefore, we consider AudioSet as a unimodal dataset and VGG

*Corresponding to hangzhao@mail.tsinghua.edu.cn



(a) UM teacher (b) NOISY student (c) MM student

Figure 1: normalized confusion matrix test accuracy

Sound as multimodal. Audios from AudioSet and audio-video pairs from VGGSound are taken as the labeled unimodal dataset D_l and unlabeled multimodal data D_u respectively. Similarly, a student network is given soft pseudo labels of the UM teacher for training.

2. Experimental Results

2.1. Emotion Recognition

One interesting finding is presented in Figure 1. We compare the confusion matrix that the UM teacher, NOISY student and our MM student generates on test data. Compared with NOISY student, the MM student contributes quite differently for 8 classes: it significantly improves the class "surprised" and slightly improves over the "neutral" class. We hypothesize that audios belonging to class "surprised" have more distinct features than "neutral", and a multi-modal student effectively utilizes this information.

2.2. Semantic Segmentation

Figure 3 presents more segmentation results on NYU Depth V2 test data. We can see that the UM Teacher generates inconsistent and noisy predictions, for instance, they fail to identify sofas in the third, fourth and sixth example. NOISY Student improves a little over the teacher's prediction. However, its prediction is still messy. In contrast, MM student identifies the sofa as a whole and gives mostly correct predictions. Depth modality here enables knowledge expansion from the RGB teacher.

2.3. Event Classification

We list top 5 event categories that our MM student improves most in Table 1. While NOISY student leads to similar performance gain for each event class, our MM student greatly improves over these classes with the assistance of video modality. For instance, the UM teacher performs poorly on the "dog growling" class with audio inputs only. NOISY student improves test mAP from 0.069 to 0.096 with the help of more data. In contrast, a MM student achieves an mAP of 0.542 and shows great improvement over the unimodal baselines. Video modality helps our MM student denoise these incorrect predictions given by the UM teacher.

	Test mAP		
	UM teacher	NOISY student	MM student (ours)
basketball bounce	0.178	0.263	0.542
dog growling	0.069	0.096	0.516
people belly laughing	0.334	0.475	0.800
sliding door	0.104	0.163	0.388
lawn mowing	0.318	0.481	0.541

Table 1: Performance of top 5 event categories that MM student improves. Test mAP of the UM teacher and NOISY student are shown for comparison.

3. Ablation Studies

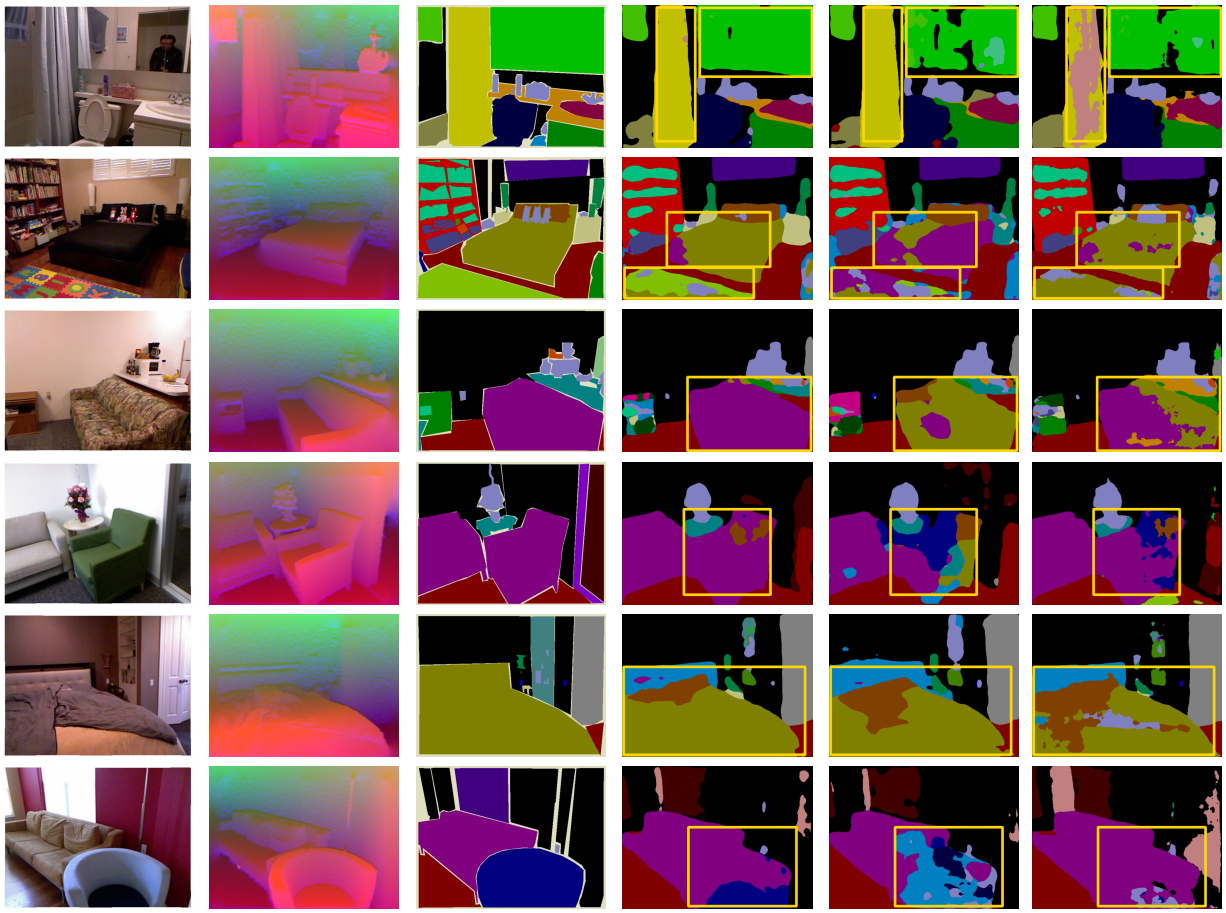
In this section, we provide a comprehensive study of various factors in *MKE*.

3.1. Regularization

The ablation study for regularization terms is provided in the main paper. We report performance of MM student (no reg), *i.e.*, a MM student without regularization in all experiments. Results consistently show that a MM student yields better results than a MM student (no reg). We arrive at the conclusion that multimodality combined with regularization leads to best performance compared with all the baselines.

3.2. Unlabeled Data Size

We study the effect of unlabeled data size in this section. Specifically, for the task of semantic segmentation, we reduce unlabeled data size from 1488 RGB-D image pairs as reported in the main paper to 744 image pairs. Results are shown in Table 2.



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

Figure 3: Qualitative segmentation results on NYU Depth V2 test set.

Method	Train data			Test mIoU (%)
	mod	D_l	$\tilde{D}_u$	
UM teacher	rgb	✓		44.15
UM student	rgb		✓	44.57
NOISY student	rgb	✓	✓	46.85
MM student (ours)	rgb, d		✓	**47.44**

Table 2: Results of semantic segmentation on NYU Depth V2. We set unlabeled data size smaller than labeled data size.

UM student yields marginal improvement over UM teacher as it receives a small amount of unlabeled data and pseudo labels for training. On the contrary, provided with same data as the UM student, a MM student still achieves a mIoU gain of 3.29%. Furthermore, although training data of NOISY student is twice greater than that of a MM student, half of which contain true labels, our MM student still achieves better results with respect to NOISY student. The great denoising capability of *MKE* is thus shown.

3.3. Teacher Model

The UM teacher of previous experiments on NYU Depth V2 is implemented as DeepLab V3+. In this section, we experiment with the teacher model as RefineNet. We utilize same data as in Section 4.2, where $|D_l| = 795$, $|D_u| = 744$, and $|D_{test}| = 654$. Table 3 reports performance when the UM teacher is RefineNet with ResNet-50 and ResNet-101 as backbone respectively.

Method	*mod*	Test mIoU(%)	
		RefineNet-Res50	RefineNet-Res101
UM teacher	rgb	42.41	44.18
UM student	rgb	41.23	42.89
NOISY student	rgb	43.21	45.69
MM student	rgb, d	**45.71**	**46.95**

Table 3: Ablation study for UM teacher model architecture. MM student consistently denoises pseudo labels when teacher model varies.

Despite different model architectures of the UM teacher, the conclusion holds same: MM student significantly outperforms the UM teacher and UM student, achieving knowledge expansion. In addition, a stronger teacher (*i.e.*, more reliable pseudo labels) will lead to a better student model in the case of both unimodality and multimodality. Another observation here is that UM student fails to surpass UM teacher due to limited size of D_u. On the contrary, given small amount of unlabeled data, our MM student effectively utilizes unlabeled multimodal data and outperforms NOISY student which has access to both labeled and unlabeled data.

3.4. Pseudo Labels for Distilling

We also investigate how soft and hard pseudo labels influence results and report results in Table 4. We follow same data and model settings in the previous section.

As shown in Table 4, soft labels yield slightly better results than hard labels. The MM student learning from soft labels of the UM teacher achieves highest test mIoU.

Method	*mod*	Labels for distilling	Test mIoU(%)
UM teacher	rgb	$\star$	44.18
UM student	rgb	hard	42.53
UM student	rgb	soft	42.89
MM student	rgb, d	hard	46.64
MM student	rgb, d	soft	**46.95**

Table 4: Ablation study for hard *vs.* soft labels on semantic segmentation. $\star$ means that the UM teacher is trained on true labels. Other methods are trained on pseudo labels generated by the UM teacher.

4. Proofs

4.1. Equivalence of Loss Terms

We prove below that Equation (3) is equivalent to Equation (3) in the main paper.

$$\theta_s^\star = \underset{\theta_s}{\operatorname{argmin}} \frac{1}{M} \sum_{i=1}^{M} l_{cls}(\tilde{\mathbf{y}}_i, \mathcal{T}(\mathbf{f}_s(\mathbf{x}_i^\alpha, \mathbf{x}_i^\beta; \theta_s)) \quad (3)$$

l_{cls} refers to cross entropy loss for hard labels and KL divergence loss for soft labels. It takes the form of:

$$l_{cls}(y, p) = -\sum_{k=1}^{K} y_k \log \frac{\exp p_k}{\sum_{j=1}^{K} \exp p_j} + \sum_{k=1}^{K} y_k \log y_k \quad (4)$$

where y and p are K-dimensional vectors. K denotes the number of classes. For simplicity, let z denote the output of feeding p into a softmax layer, *i.e.*, $\forall k \in [K], z_k = \frac{\exp p_k}{\sum_{j=1}^{K} \exp p_j}$.

The derivative of $l_{cls}(y, p)$ with respect to p_j is:

$$\begin{aligned} \frac{\partial l_{cls}(y,p)}{\partial p_j} &= -\sum_{k=1}^{K} y_k \frac{\partial \log z_k}{\partial p_j} \\ &= -\sum_{k=1}^{K} y_k (I_{kj} - z_j) = z_j - y_j \end{aligned} \quad (5)$$

Therefore, $\nabla l_{cls} = [z_1 - y_1, z_2 - y_2, ..., z_k - y_k]$.

$$||\nabla l_{cls}|| = \sqrt{\sum_{j=1}^{K} (y_j - z_j)^2} \leq \sqrt{K} \quad (6)$$

Equation (6) states that $l_{cls}(y, p)$ is Lipschitz continuous in p for fixed y with respect to $||\cdot||$, where $\sqrt{K}$ is the Lipschitz constant. Therefore, $\exists -\sqrt{K} \leq \gamma \leq \sqrt{K}$, such that loss terms in Equation (3) equal to that of Equation (3) in the main paper.

4.2. Lemma 1

To start with, by definition of (a, c) expansion and $max(c_1, c_2) \leq \frac{1}{\bar{a}}$, we derive Equation (7) and (8) from Equation (10) and (11) in the main paper.

$$\begin{aligned} &P_i(N(V^\alpha)) \geq c_1 P_i(V^\alpha) \\ &\forall\, V^\alpha \subseteq \mathcal{X}^\alpha \; with \; P_i(V^\alpha) \leq \bar{a} \end{aligned} \quad (7)$$

$$\begin{aligned} &P_i(N(V^\beta)) \geq c_2 P_i(V^\beta) \\ &\forall\, V^\beta \subseteq \mathcal{X}^\beta \; with \; P_i(V^\beta) \leq \bar{a} \end{aligned} \quad (8)$$

Multiplying both sides of Equation (7) and Equation (8), we have:

$$\begin{aligned} &P_i(N(V^\alpha)) P_i(N(V^\beta)) \geq c_1 c_2 P_i(V^\alpha) P_i(V^\beta) \\ &\forall\, V^\alpha \subseteq \mathcal{X}^\alpha \; with \; P_i(V^\alpha) \leq \bar{a} \\ &\forall\, V^\beta \subseteq \mathcal{X}^\beta \; with \; P_i(V^\beta) \leq \bar{a} \end{aligned} \quad (9)$$

Plugging in conditional independence (*i.e.*, Equation (12) in the main paper) gives us:

$$\begin{aligned} &P_i(N(V)) \geq c_1 c_2 P_i(V), \\ &\forall\, V \subseteq \mathcal{X} \; with \; P_i(V) \leq \bar{a} \end{aligned} \quad (10)$$

Thus, P on $\mathcal{X}$ satisfies $(\bar{a}, c_1 c_2)$ expansion.