Missing Modality Robustness in Semi-Supervised Multi-Modal Semantic Segmentation

Harsh Maheshwari[1,2*] Yen-Cheng Liu[1] Zsolt Kira[1]
[1]Georgia Institute of Technology, [2]Avataar.ai
{harshmaheshwari, ycliu, zkira}@gatech.edu

Abstract

Using multiple spatial modalities has been proven helpful in improving semantic segmentation performance. However, there are several real-world challenges that have yet to be addressed: (a) improving label efficiency and (b) enhancing robustness in realistic scenarios where modalities are missing at the test time. To address these challenges, we first propose a simple yet efficient multi-modal fusion mechanism Linear Fusion*, that performs better than the state-of-the-art multi-modal models even with limited supervision. Second, we propose* M3L*:* M*ulti-modal Teacher for* M*asked* M*odality* L*earning, a semi-supervised framework that not only improves the multi-modal performance but also makes the model robust to the realistic missing modality scenario using unlabeled data. We create the first benchmark for semi-supervised multi-modal semantic segmentation and also report the robustness to missing modalities. Our proposal shows an absolute improvement of up to* 5% *on robust mIoU above the most competitive baselines. Our project page is at https://harshm121.github.io/projects/m3l.html*

1. Introduction

The availability of multiple sensors such as RGB, depth, and infrared has encouraged the use of multiple modalities for scene understanding tasks like semantic segmentation [4, 19, 20, 39, 42–44]. Multi-modal semantic segmentation has shown promising results outperforming their unimodal counterparts [7,21,48] due to the effective use of auxiliary information present across modalities. However, one challenge with semantic segmentation is getting substantial amounts of annotated data, which is a laborious and costly process. This has encouraged a need to create algorithms that work well with limited supervision. One approach is to utilize a mixture of labeled and unlabeled data (*i.e.,* semi-supervised learning), and several works have approached this for various tasks [3,24,25,46]. To the best of our knowledge, however, all of the semi-supervised semantic segmentation research has been focused on uni-modal segmentation [8, 16, 28, 45, 52, 53]. Thus, there is a need to explore semi-supervised frameworks for multi-modal semantic segmentation that can effectively use additional modalities to make the task label efficient.

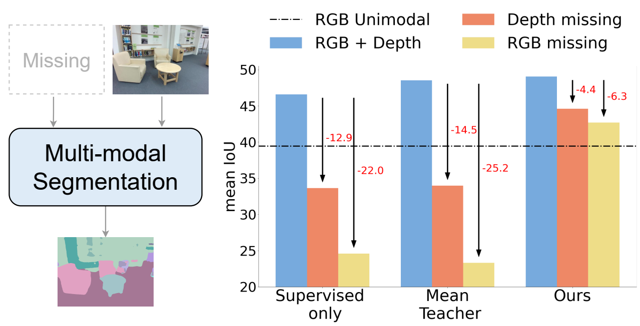

Figure 1. Existing multi-modal semantic segmentation models require a significant amount of training data and exhibit lower robustness when dealing with missing modalities during test time. Our proposed framework can learn the model with a limited amount of labeled data and enhance the robustness to test-time missing modality scenarios.

We find that there are two major challenges for making multi-modal segmentation models more useful that need attention. The first is to create a modality fusion algorithm that can work well even with limited supervision. The current multi-modal literature [22,42] has focused on fully supervised scenarios and thus the resulting methods do not necessarily work well with limited supervision. There has also been a growing interest in using transformers for multiple modalities due to their flexibility to incorporate various data types. The state-of-the-art multi-modal semantic segmentation method [42] uses a learned fusion mechanism with a transformer-based segmentation architecture, Segformer [48]. The use of transformers and training additional parameters to learn the fusion mechanism have made it more challenging for the existing models to perform well in a low-label regime.

The second challenge is a lack of robustness to test-time

[*]work done while at Georgia Institute of Technology

missing modalities. Multi-modal models show an improvement over their uni-modal counterparts by effectively fusing the auxiliary information from different modalities. However, this improvement comes with a stricter requirement of guaranteeing that all of the modalities will be present during test time. This requirement could be difficult to satisfy due to sensor failures and unreliability. As previously discussed in the medical domain [10, 18, 31], we also discover a major weakness with current multi-modal semantic segmentation - missing modality robustness. We find that if any modality is missing during test time, the segmentation performance of such models falls drastically, even below their uni-modal counterparts as depicted in Figure 1. Learning with limited supervision and robustness to missing modalities are realistic and practical challenges in the existing semantic segmentation works. These may co-exist as getting segmentation labels is a costly process, and relying on additional modalities to improve performance makes the model prone to severe degradation when any modality is missing. Importantly, the two interact - we found that a naive combination of the solutions to the two problems (semi-supervised - mean teacher and missing modality - modality dropout augmentation) cannot address this problem.

To address the above limitations, we introduce *Linear Fusion*, a multi-modal segmentation model that works effectively even with limited supervision, and *M3L*: *M*ulti-modal Teacher for *M*asked *M*odality *L*earning, a semi-supervised framework which effectively uses unlabeled data to not only improve multi-modal semantic segmentation performance but also to make the model robust to missing modalities.

Specifically, *Linear Fusion* is a simple yet effective multi-modal segmentation model that combines tokens from the two modalities linearly and thus learns cross-modal interaction without using additional trainable parameters. This makes the simple algorithm effective even when trained with limited supervision. For example, when trained with 0.2% data on Stanford Indoor dataset [1], Linear Fusion outperforms the current state-of-the-art by 3.5% points mIoU. In addition, to both leverage unlabeled data and enhance the robustness to missing modality, we propose *M3L*, a semi-supervised framework that trains a Linear Fusion model and uses a multi-modal mean teacher to supervise a student network with a randomly chosen modality masked in the input. This makes the model robust to missing modalities while improving segmentation performance. Surprisingly, we find that a bi-modal model trained with our framework, when given a single input, still performs better than its uni-modal semi-supervised counterparts, and thus M3L can also be used to improve uni-modal semi-supervised semantic segmentation by using privileged multiple modalities during training (details in Section 4.3.2).

We perform extensive experimentation and comparison of our proposed methods against existing baselines to verify our claims and show the effectiveness of our proposals. We show that Linear Fusion, when trained with M3L, shows an improvement of up to 5% points mIoU above the most competitive baseline on robust multi-modal segmentation. Moreover, for uni-modal segmentation, our method shows an absolute improvement of up to 3.5% mIoU for RGB uni-modal and up to 6.5% mIoU for depth uni-modal over the semi-supervised uni-modal segmentation baselines.
Finally, we list all our contributions:

1. To the best of our knowledge, we are the *first* to address semi-supervised multi-modal semantic segmentation, and we create a new benchmark and evaluate robustness to realistic test-time missing modality scenarios.

2. We present a simple yet effective cross-modal integration mechanism, Linear Fusion, which outperforms state-of-the-art [42] under limited supervision.

3. We propose a semi-supervised training framework, M3L, which utilizes unlabeled images to improve the segmentation performance and make the model robust to missing modalities.

We will release our code to encourage further research.

2. Related Work

Multi-modal semantic segmentation. As accessing multiple spatial modalities like RGB, depth, and infrared is getting easier, many methods have been proposed to use more than one modality to improve semantic segmentation performance. The holy grail of the multi-modal semantic segmentation community is to find effective ways to fuse auxiliary information from multiple modalities. These methods can be broadly categorized into early [9, 12, 35], late [51], and mid/hybrid [14, 19, 22, 39, 42, 43] fusion techniques. Convolutional models have dominated this literature so far, and interest has arisen in using transformer-based architectures for fusing multiple modalities [22, 42]. However, these works focused on creating fusion mechanisms for fully supervised multi-modal semantic segmentation, and there is no prior work on *multi*-modal semi-supervised semantic segmentation.
Semi-supervised semantic segmentation. Creating labels for segmentation is a more laborious task and can cost around $25\times$ more than getting labels for classification[1]. This has motivated a lot of research on semi-supervised semantic segmentation [8, 16, 23, 28, 45, 52, 53]. We find that most of these methods are smart extensions of the popular mean teacher framework [37] which uses a weight-ensembled teacher model to generate pseudo labels. How-

[1] https://cloud.google.com/ai-platform/data-labeling/pricing

Related Area	Multi-modal	Semi-supervised	Missing modality robustness	Related work
Multi-modal segmentation	✓			[9, 12, 14, 19, 22, 35, 39, 42, 43, 51]
Semi-supervised segmentation		✓		[8, 16, 23, 28, 45, 52, 53]
Missing modality robustness	✓		✓	[2, 10, 13, 15, 18, 30, 31, 34, 41, 47, 50]
Ours	✓	✓	✓	-

Table 1. Related literature. Missing modality robustness is unexplored for semantic segmentation. We address the two simultaneous challenges of making the multi-modal models robust to missing modality while making them label efficient, thus more practical and useful.

ever, we found that all of the semi-supervised semantic segmentation work focused on only RGB uni-modal models, and an investigation of how additional modalities can be used to more strongly leverage unlabeled data is an interesting research question. Thus, we found a rising interest in semi-supervised learning using multiple modalities in other domains such as medical [5, 6], videos [49], speech [36] but not for semantic segmentation. Our proposed approach thus uses a semi-supervised setup for multi-modal semantic segmentation and extends the mean teacher [37] framework to not only improve segmentation performance but also make the model robust to missing modalities.

Missing modality robustness. The use of multiple modalities is common in various domains and thus the robustness to missing modalities has become an important area of focus. For example, in medical domains, it is possible to not have access to all modalities (all types of scans like MRI, CT, etc.) for all patients. Thus missing modality robustness in the medical domain has caught a lot of attention. The approaches mainly include either synthesizing the missing modality [15, 30, 50], or learning a shared latent space for all modalities [10, 18, 31] or knowledge distillation methods [2, 41]. Modality dropout has also been used to improve performance in certain domains by ensuring enough attention to all modalities and as a regularization to not let the dominant modality drive the prediction. Neverova *et al.* [27] proposed ModDrop as augmentation for gesture recognition, [47] proposes ActionMAE for action recognition which is robust to missing modalities and Abdelaziz *et al.* [13] proposed modality dropout for driving animated talking faces for audio-video modalities and [34] for multi-modal dialogue systems. We show that just using the modality dropout augmentations is not enough for a semi-supervised training framework and thus M3L uses a knowledge distillation framework to utilize the unlabeled data to make the models robust to missing modality. Frameworks that gain performance by utilizing multiple modalities also become prone to failure when the modalities are not guaranteed to be present during test time. This has gained attention in other domains and we highlight the same for multi-modal semantic segmentation.

Robustness in segmentation. In this work, we focus on robustness to missing modalities at test time. Robustness in segmentation also can be to other forms of degradation. Tian *et al.* [38] presents a method to fuse multiple modalities for segmentation effectively when certain modalities may suffer from degradations like motion blur, gaussian noise, fog, etc. Robustness to seasonal changes or lighting effects has also been discussed in prior works [17, 29, 40]. However, our work focuses on a scenario when the entire modality is missing, which is another possible situation due to sensor malfunction or other unreliabilities.

We summarise the related work in Table 1. We focus on a problem setting that has been explored in parts. There has been a growing interest in multi-modal segmentation and semi-supervised segmentation. We also find a growing interest in missing modality robustness in other domains but it has not been discussed for semantic segmentation. Thus, to make segmentation models more useful and practical, there is a need to address both, the label-efficiency and the robustness to missing modalities of such models. We thus propose our method M3L to address both challenges.

3. Method

Problem definition. Our ultimate goal is to address the missing modality robustness of multi-modal segmentation models when trained with limited supervision. We consider our data has two modalities for training: RGB, denoted by x^{rgb}, and depth, denoted by x^{depth}. The goal is to output a segmentation map classifying each pixel into one of $\mathbb{C}$ classes. We consider a set of labeled samples, $\mathcal{D}_s = \{x_i^{rgb}, x_i^{depth}, y_i\}_{i=1}^{N_s}$ and a set of unlabeled samples $\mathcal{D}_u = \{x_i^{rgb}, x_i^{depth}\}_{i=1}^{N_u}$. N_s and N_u are the numbers of labeled and unlabeled data. To examine the performance and robustness of any algorithm, say $\mathcal{A}$, we report the performance P of the predictions under three test conditions: a) Using both the modalities: $P(\mathcal{A}(x^{rgb}, x^{depth}), y)$, b) with RGB only: $P(\mathcal{A}(x^{rgb}), y)$ and c) with depth only: $P(\mathcal{A}(x^{depth}), y)$.

To address missing modality robustness, we propose M3L, a semi-supervised training framework for making the multi-modal semantic segmentation models robust to missing modalities (Section 3.3). We do so by devising a knowledge distillation framework, leveraging pseudo labels from a multi-modal teacher to a masked modality student model. However, before talking about the semi-supervised framework, we pay attention to the base multi-modal model and find that the existing state-of-the-art fusion mechanism [42] does not perform well in a low-label regime (Section 3.1). Hence, we first devise Linear Fusion, a simple yet effective fusion mechanism for multi-modal semantic segmentation (Section 3.2), and then train the Linear Fusion model using M3L, to improve the segmentation performance and make the model robust to missing modalities.

3.1. Revisiting Multi-modal Semantic Segmentation

With the goal of improving semantic segmentation using auxiliary information from different spatial modalities, Wang et al. proposed TokenFusion [42]. It is a multi-modal transformer architecture that dynamically detects uninformative tokens from a modality and substitutes them with tokens from other modalities allowing the transformer to learn cross-modal interactions. The detection of uninformative tokens is achieved by thresholding a score estimated by a separate scoring module. Unlike Segformer [48] (Eq. 1), TokenFusion (Eq. 2) uses the score from the scoring module as weights for each token before passing it to the next layer. The substitution is done as shown in Eq. 3 and is encouraged by making the scores sparse using **L1** loss on the scores to learn cross-modal interactions.

$$\hat{e}^l_m = \text{MHA}(\text{LN}(e^l_m)), e^{l+1}_m = \text{FF}(\text{LN}(\hat{e}^l_m)) \quad (1)$$

$$\hat{e}^l_m = \text{MHA}\big(\text{LN}(e^l_m) \cdot s^l(e^l_m)\big), e^{l+1}_m = \text{FF}(\text{LN}(\hat{e}^l_m)) \quad (2)$$

$$e^l_m = e^l_m \odot \mathbb{I}_{s^l(e^l_m)\geq\theta} + e^l_{m'} \odot \mathbb{I}_{s^l(e^l_m)<\theta} \quad (3)$$

Here, e^l_m denotes the tokens out of the l^{th} layer for the m^{th} modality, and θ is the exchange threshold. MHA, LN, and FF are Multi-Headed Attention, Layer Normalization, and Feed Forward modules as described in Segformer [48].

Limitation. We find that this learned substitution mechanism does not work very well (more details in Section 4.3.3) when trained with small amounts of data. To this end, we propose a much simpler fusion mechanism that has significant performance improvement over Token Fusion.

3.2. Linear Fusion for Cross-Modal Integration

Fusing information from multiple modalities is the holy grail of multi-modal segmentation. Multiple methods have been presented in the literature for convolution models [19, 20, 39, 43, 44] but only recently has the interest arisen to use transformer models for using multiple modalities for segmentation. However, it is a challenging task with limited supervision where there is not enough data to train complex learning-based fusion mechanisms.

Base Model. We consider Segformer [48], as our base segmentation model based on the transformer architecture due to its popularity in the vision community. To extend the uni-modal Segformer architecture to handle multiple input modalities, we create two copies of the hierarchical transformer encoder f from Segformer which share weights θ.

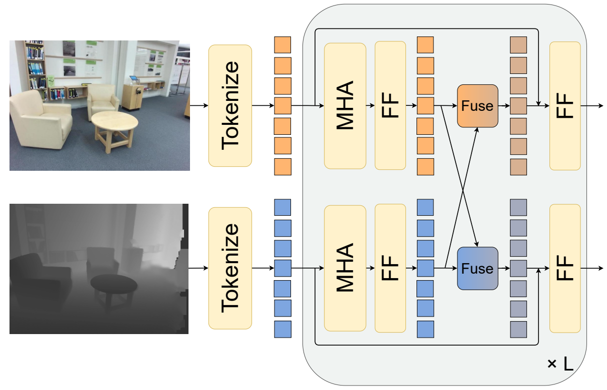

Figure 2. Overview of the Linear Fusion model. Information from the two modalities is fused by linearly adding the tokens from each branch. The fused tokens are then passed to further layers. MHA denotes Multi-Headed Attention and FF denotes a feed-forward module. Tokenize, MHA, and FF are the same as in the Segformer [48] architecture

However, as proposed in Token Fusion [42], we use separate layer normalization parameters γ for the two modalities as the statistics of different modalities can be vastly different. Each branch gets a single modality x^m as input and the final representation of each branch is passed to a lightweight MLP decoder g with parameters ϕ, as proposed in Segformer. Specifically,

$$\hat{y}^m = g_\phi \circ f_{\theta,\gamma^m}(x^m, \mathbf{e}^{\overline{m}}), \quad (4)$$

where $\hat{y}^m$ is the prediction from a branch corresponding to the modality m. Note that the encoder f takes $\mathbf{e}^{\overline{m}}$ as an additional argument, which is a vector of all the tokens from intermediate layers of the other branch $\overline{m}$ and is used for fusion, as described below.

Fusion. We apply a simple fusion mechanism that integrates cross-modal tokens by using a linear combination. Specifically, each of the two branches gets a single modality x^m as input to the encoder f. The encoder f has L attention layers and the intermediate token outputs from each attention layer e_l are fused together,

$$e^l_m = \alpha \times e^l_m + (1-\alpha) \times e^l_{\overline{m}}, \quad (5)$$

where the hyperparameter α is the fusion weight guiding the extent of information preserved in the branch. In this way, each branch combines the information it receives from the other branch by linearly combining the tokens. We depict this fusion mechanism in Figure 2. $\mathbf{e}^m$ represented in Eq. 4 is a vector of $[e^l_m]^L_{l=1}$.

Ensemble prediction. As shown in Eq. 4, each branch outputs an individual prediction driven by input from a modality and fusion of information from the other. The final prediction of the model is a weighted ensemble of $\hat{y}^m \ \forall \ m$.

$$\hat{y} = \lambda\hat{y}^{rgb} + (1-\lambda)\hat{y}^{depth} \quad (6)$$

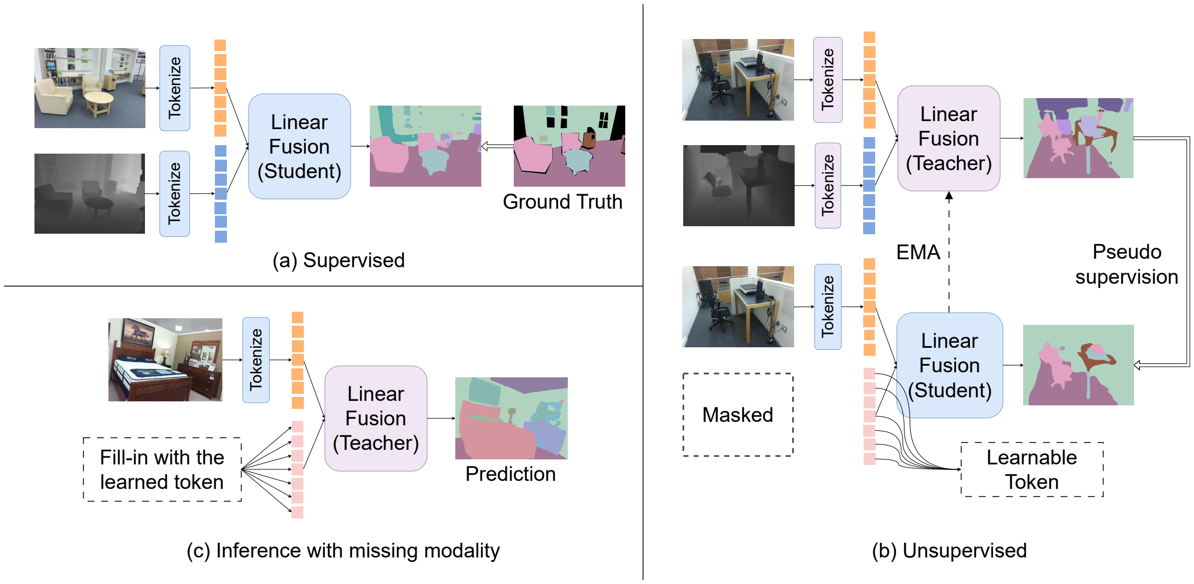

Figure 3. Overview of the M3L framework. (a) M3L supervises the prediction using ground truth for the labeled instances. (b) For the unsupervised loss, M3L uses a multi-modal mean (EMA) teacher which generates a segmentation prediction that is used to supervise a student. A randomly chosen modality is masked entirely in the student's input and a single learnable token is used to fill in the missing tokens. (c) The learned token thus can be used during inference if any modality is missing.

where λ is a trainable ensemble weight. Essentially, the model outputs three predictions $\hat{y}^{rgb}$, $\hat{y}^{depth}$ from the two branches and the ensemble $\hat{y}$. The model is trained using supervised loss $\mathcal{L}_s$ which is the average of segmentation loss $\mathcal{L}_{seg}$ of all three predictions with the ground truth y,

$$\mathcal{L}_s = \frac{1}{3}\left[\mathcal{L}_{seg}(\hat{y}, y)+\mathcal{L}_{seg}(\hat{y}^{rgb}, y)+\mathcal{L}_{seg}(\hat{y}^{depth}, y)\right] \quad (7)$$

The above method describes an effective way to integrate the cross-modal features, and our empirical results in Section 4.3.3 indicate our simple linear fusion performs favorably against the prior work [42] under the low-label (and even the full-labeled) settings. With this base segmentation model, we now proceed towards a semi-supervised framework to help with missing modality robustness for multi-modal segmentation.

3.3. M3L: Multi-modal Teacher for Masked Modality Learning

To further improve label efficiency, in this section we propose a training framework that leverages unlabeled data to improve the performance of the multi-modal segmentation model and makes the model robust to missing modalities. We consider the base segmentation model to be Linear Fusion and denote it by LF which outputs a segmentation prediction $\hat{y} = LF(x^{rgb}, x^{depth})$. As presented in Figure 3, we describe our M3L framework as follows.

Supervised training. For the labeled samples in $\mathcal{D}_s$, we compute the segmentation loss described in Eq. 7 to train the base segmentation model, as depicted in Figure 3 (a).

Unsupervised training. To leverage the unlabeled data for improving the performance and missing modality robustness, we propose a semi-supervised framework based on a teacher-student mechanism [37] and a modality dropout scheme as shown in Figure 3 (b). Specifically, the framework consists of a teacher (LF^t) and a student (LF^s) model which are identical network architectures but do not share weights[2]. Our teacher model, LF^t takes both the modalities (RGB **and** depth) as input and estimates the segmentation mask using Eq. 4 and 6.

$$\hat{y}_t = LF^t(x^{rgb}, x^{depth}) \quad (8)$$

We then generate *hard* pseudo-labels, $y_p = \arg\max\ \hat{y}_t$. To improve the label-efficacy of the student model and to make it more robust to the missing modality, we propose to randomly mask 100% of either modality (RGB **or** depth) in the input to the student model. To handle the missing modality, we use a learnable token to fill in for all the missing tokens of the masked modality.The learned token can thus be used during inference whenever any modality is missing, as shown in Figure 3 (c). Modality dropout makes the student model robust to missing modalities by not only providing a learned token to fill-in whenever needed but

[2]We overload the notation for simplicity and use θ_t and θ_s to represent the teacher and student's parameters.

also by encouraging the model to pay attention to all modalities and discouraging the dominating modality to overpower. The student predictions are supervised using the *hard* pseudo-labels generated by the teacher.

$$\mathcal{L}_u = \frac{1}{3}\Big[\mathcal{L}_{useg}(\hat{y}_s, y_p) + \mathcal{L}_{useg}(\hat{y}_s^{rgb}, y_p) + \mathcal{L}_{useg}(\hat{y}_s^{depth}, y_p)\Big] \tag{9}$$

where $\mathcal{L}_{useg}$ is the unsupervised segmentation loss. The loss $\mathcal{L}_u$ is computed on both the labeled and unlabeled samples.

Overall loss. We train the overall framework using a batch of both labeled and unlabeled samples. The supervised loss L_s (Eq. 7) is computed on the labeled samples and the unsupervised loss L_u (Eq. 9) is computed on both the labeled and unlabelled samples.

$$\mathcal{L} = \sum_{(\mathbf{x},y)\sim\mathcal{D}_s} \mathcal{L}_s(\mathbf{x}, y) + \lambda_{pseudo} \sum_{(\mathbf{x},y)\sim\mathcal{D}_s\cup\mathcal{D}_u} \mathcal{L}_u(\mathbf{x}) \tag{10}$$

We train the student model by backpropagating the total loss $\mathcal{L}$ but detach the teacher parameters from the computation graph. The teacher model's parameters are updated slowly using the student model's parameters.

Teacher update. To obtain stable pseudo-labels from the teacher model, we set the teacher model's parameters as the exponential moving average (EMA) of the student's parameters. The slowly progressing teacher can be regarded as a temporal ensemble of the student model across training iterations.

$$\theta_t \leftarrow \alpha_{ema}\theta_t + (1 - \alpha_{ema})\theta_s \tag{11}$$

with α_{ema} being the hyperparameter controlling the rate of update of the teacher parameters. EMA teacher proposed in [37] has been successfully used in various other tasks [3, 23, 24, 45].

4. Experiments

To demonstrate the effectiveness of Linear Fusion (Section 3.2) and M3L (Section 3.3), we show empirical evidence by comparing against many baselines and ablate over our design choices.

4.1. Benchmark for Semi-supervised Multi-modal Segmentation

Since no prior work explores semi-supervised multi-modal semantic segmentation, we create a *new* benchmark for this setting[3]. We consider the challenging scenarios of missing modality (RGB-only or depth-only) and evaluate the robustness in the proposed and existing multi-modal semantic segmentation models.

Datasets. Following prior multi-modal semantic segmentation literature [39, 42], we build the benchmark using indoor RGBD segmentation datasets. We use two popular semantic segmentation datasets - Stanford Indoor [1] and SUN RGBD [33]. We apply a more rigorous experiment setup, which splits the training into training and validation and keeps the testing set intact[4]. We hope this encourages better ML practices by tuning the hyperparameter on only the validation set. We then create labeled/unlabeled training sets of varying sizes for semi-supervised setup.

Stanford-Indoor [1] is a large-scale dataset with 17593 test samples and (originally) 52903 train samples. All the samples correspond to 6 areas (areas 1-6) with area 5 used for the test set [1]. We split the original 52903 training samples corresponding to the other 5 areas into train/val sets of sizes (49199 / 3704) with area 3 being used for validation. The dataset has 13 classes for the classification of pixels. We create three different semi-supervised configurations using 0.1% (49), 0.2% (98), and 1% (491) of the data as labeled and the rest as unlabeled (49159, 49101, 48708).

SUN RGBD is a challenging dataset with 37 classes and has 5049 test samples. It originally had 5285 training samples which we split into 90/10 train/val (4757 / 528) samples. For semi-supervised training, we create three configurations and treat 6.25% (297), 12.5% (594), and 25% (1189) of the data as labeled and the rest as unlabeled (4460, 4163, 3568).

Baselines. We compare our proposed M3L semi-supervised framework against competitive baselines to prove its efficacy. We also compare the performance of Linear Fusion with supervised-only transformer-based fusion approaches presented in the literature [18, 42].

- URN [18]: We choose a straightforward fusion mechanism that closely resembles the approach presented in Unified Representation Network [18], but we use Segformer as the base architecture. Different encoders are used for the two modalities and the averaged representations are passed to a single decoder. Since 2 encoders are trained, URN has $\sim 2\times$ the number of parameters.
- TF [42]: The state-of-the-art Token Fusion [42] framework. We train the model using our setup to report the performance.
- LF: Linear Fusion method as proposed in Section 3.2.
- LF + MT [37]: We train Linear Fusion with the semi-supervised mean-teacher [37] framework and report its performance.
- {method} + MD: We train the above-mentioned methods with the naive modality dropout augmentation[5].
- LF + M3L (ours): We train Linear Fusion with the proposed M3L semi-supervised framework as proposed in Section 3.3.

Metrics. For evaluation, we do *single-scale*, *non-sliding* testing by rescaling the input images to the expected model-input size and rescaling the predictions back to the original ground truth size using bilinear interpolation as done before [8, 45]. We report the mean IoU of all the methods when tested under 3 types of inputs: RGB+depth, RGB-only (depth missing), and depth-only (RGB missing). We report the average of the three scenarios as the *MM-Robust* performance to quantify the missing modality robustness of multi-modal semantic segmentation. We also report the mean class accuracy and pixel accuracy in the appendix.

4.2. Implementation Details

For all our models, we use the transformer-based model MiT-B2 proposed by Segformer [48] for a fair comparison. We initialize the network with ImageNet-1k pre-trained checkpoint available publicly [48]. For Linear Fusion, we tuned our fusion weight

[3]To the best of our knowledge, we are the first to create a benchmark for semi-supervised multi-modal semantic segmentation.

[4]We find the previous benchmark for multi-modal semantic segmentation [4, 39, 42, 43] do not have validation set, which is possible to suffer from the overfitting issue on the test set.

[5]Whenever a modality is missing during training, we fill the missing tokens with a learnable token proposed in Section 3.3.

Method	0.1% (49)				0.2% (98)				1% (491)			
	RGB	Depth	RGBD	MM-Robust	RGB	Depth	RGBD	MM-Robust	RGB	Depth	RGBD	MM-Robust
Uni-modal [48]	35.43	34.05	-	-	39.45	35.24	-	-	46.45	44.78	-	-
TF [42]	29.96	29.98	40.17	33.37	33.11	31.47	43.04	35.87	37.34	28.33	51.85	39.17
URN [18]	30.56	25.85	40.17	32.19	35.71	25.14	45.87	35.57	36.25	33.27	52.07	40.53
LF	33.96	25.09	42.09	33.71	33.66	24.61	46.60	34.96	33.51	23.70	52.47	36.56
LF + MT [37]	32.37	22.92	41.77	32.35	33.99	23.34	48.54	35.29	33.65	22.42	54.32	36.80
LF + MT + MD	35.11	34.69	39.06	36.29	41.78	39.52	46.81	42.70	48.22	46.12	53.97	49.44
Ours	**40.05** (+4.94)	**39.93** (+5.24)	**44.10** (+2.01)	**41.36** (+5.07)	**44.62** (+2.84)	**42.70** (+3.18)	**49.05** (+0.51)	**45.46** (+2.76)	**49.28** (+1.06)	**46.79**(+0.67)	**55.48** (+1.16)	**50.52** (+1.08)

(a) Stanford Indoor dataset

Method	6.25% (297)				12.5% (594)				25% (1189)			
	RGB	Depth	RGBD	MM-Robust	RGB	Depth	RGBD	MM-Robust	RGB	Depth	RGBD	MM-Robust
Uni-modal [48]	28.71	22.81	-	-	35.33	27.6	-	-	38.31	30.43	-	-
TF [42]	27.97	23.58	29.31	26.95	33.75	28.31	35.88	32.65	37.36	31.90	39.86	36.37
URN [18]	28.72	12.47	31.31	24.17	33.66	15.62	37.62	28.97	37.49	17.27	40.49	31.75
LF	29.69	15.75	**32.00**	25.81	35.48	17.46	39.00	30.65	39.15	17.66	42.09	32.97
LF + MT [37]	29.57	17.86	31.11	26.18	34.82	18.89	39.17	30.96	38.96	21.03	41.95	33.98
LF + MT + MD	29.31	24.71	30.36	28.13	36.65	30.53	38.35	35.18	40.30	33.70	42.34	38.78
Ours	**29.92** (+0.23)	**25.44** (+0.73)	30.67 (-1.33)	**28.68** (+0.55)	**38.12** (+1.47)	**32.29** (+1.76)	**39.70** (+0.53)	**36.70** (+0.89)	**41.31** (+1.01)	**34.11** (+0.41)	**42.69** (+0.34)	**39.37** (+0.59)

(b) SUN RGBD dataset

Table 2. **Missing Modality robustness**. We compare the multi-modal models on three testing scenarios: RGBD, RGB (Depth missing), and Depth (RGB missing). We also report the individual uni-modal model's performance for the two modalities for comparison.

Method	0.1% (49)	0.2% (98)	1% (491)
Uni-modal (sup-only)	35.43	39.45	46.45
Uni-modal + MT [37]	36.59	41.50	47.04
Uni-modal + CPS-Seg[6] [8]	37.09	42.75	46.37
Ours	**40.05** (+2.96)	**44.62** (+1.87)	**49.28** (+2.24)

(a) RGB uni-modal

Method	0.1% (49)	0.2% (98)	1% (491)
Uni-modal (sup-only)	34.05	35.24	44.78
Uni-modal + MT [37]	33.46	37.57	46.25
Uni-modal + CPS-Seg[6] [8]	33.56	36.71	45.71
Ours	**39.93** (+5.88)	**42.70** (+5.13)	**46.79** (+0.54)

(b) Depth uni-modal

Table 3. **Multi-modal training benefits the uni-modal segmentation results.** Performance with uni-modal input of our proposed LF+M3L method as compared to uni-modal supervised only and semi-supervised (MT, CPS) models on Stanford Indoor dataset [1]

on the validation set and chose $\alpha = 0.8$ for all settings. For M3L, we either give RGB and Depth, or RGB-only, or Depth-only input to the student model (in equal proportions). We set $\alpha_{ema} = 0.99$ and $\lambda_{pseudo} = 1.0$. We use AdamW optimizer [26] and train on a minibatch of 16 with a learning rate of $1e-4$ for the encoder and $3e-4$ for the decoder with a momentum of 0.9, a weight decay of $1e-4$ and a polynomial decay of power 0.9. For semi-supervised methods, we sample 16 labeled and 16 unlabeled data instances in each minibatch. We use OHEM loss [32] as $\mathcal{L}_{seg}$ in Eq. 7 and the multi-class cross-entropy loss as $\mathcal{L}_{useg}$ in Eq. 9. We train for $\sim$ 15k iterations and pick the checkpoint (sampled after every 300 iterations) with the best validation performance to report the test performance. Any hyperparameter tuning was done on the validation set keeping the test set untouched. Addtional implementation details are specified in the appendix A.4

4.3. Results

To prove the efficacy of M3L and Linear Fusion, we perform extensive experiments on Stanford Indoor and SUN RGBD datasets and three labeled/unlabeled configurations per dataset.

4.3.1 Semi-supervised Multi-modal Segmenation

To show that M3L is effectively using unlabeled images, we compare the Linear Fusion model trained with M3L against the supervised-only baseline, a semi-supervised mean teacher [37], and mean teacher when trained with modality dropout augmentation. When tested with multi-modal RGBD input, our proposed method consistently performs better and gives an absolute improvement of up to 2.01% mIoU over the strong MT [37] baseline on Stanford Indoor [1] dataset, as shown in Table 2 (a).

To show M3L's ability to make the model robust to missing modalities, we also test the models on a more challenging scenario of missing modalities and report the MM-Robust metric, which is the average of the three possible test-time scenarios (missing depth, missing RGB, RGBD). As shown in Table 2 (a), on MM-Robust metric, M3L is better than the MT baseline by up to 5.07% mIoU and consistently shows improvement for both depth and RGB missing scenarios on Stanford Indoor dataset. We show similar results on the SUN RGBD dataset in Table 2 (b). Notably, we see that when trained with modality dropout augmentation in a semi-supervised setting, even though there is increased robustness to missing modalities, the RGBD performance drops, denoting that modality dropout augmentation is not able to effectively use additional modalities. M3L not only improves the multi-modal segmentation performance but also makes the model robust to missing modalities by effectively using the unlabeled data and additional modalities. We note that on SUN RGBD dataset, even though M3L successfully improves the performance for missing modality scenarios, neither MT [37] nor M3L sufficiently improves the multi-modal (RGBD) performance. We attribute this to a lack of a large unlabeled set, which is even more essential for a challenging dataset with 37 fine-grained classes.

Method	Trained parameters	Inference time(ms)	Stanford Indoor 0.1% (49)	0.2% (98)	1% (491)	100% (49k)	SUN RGBD 6.25% (297)	12.5% (594)	25% (1189)	100% (4.7k)
Uni-modal RGB	24.73M	17.2	35.43	39.45	46.45	50.82	28.71	35.33	38.31	45.89
Uni-modal Depth	24.73M	17.2	34.05	35.24	44.78	52.65	22.81	27.60	30.43	36.93
TF [42] (RGBD)	26.02M	55.7	40.17	43.04	<u>51.85</u>	56.64	29.31	35.88	39.86	47.00
URN [18] (RGBD)	48.93M	36.6	40.17	<u>45.87</u>	<u>52.07</u>	56.67	<u>31.31</u>	37.62	40.49	<u>47.99</u>
LF (Ours) (RGBD)	24.75M	31.7	**42.09** (+1.92)	**46.60** (+0.73)	**52.47** (+0.4)	**57.16** (+0.49)	**32.00** (+0.69)	**39.00** (+1.38)	**42.09** (+1.6)	**48.17** (+0.18)

Table 4. Comparison of supervised-only multi-modal models trained with varying amounts of data.

Method	Modality Dropout	Unlabeled	M3L	0.1% (49) RGB	Depth	RGBD	MM-Robust	0.2% (98) RGB	Depth	RGBD	MM-Robust	1% (491) RGB	Depth	RGBD	MM-Robust
TF [42]	-	-	-	29.96	29.98	40.17	33.37	33.11	31.47	43.04	35.87	37.34	28.33	51.85	39.17
URN [18]	-	-	-	30.56	25.85	40.17	32.19	35.71	25.14	45.87	35.57	36.25	33.27	52.07	40.53
TF+MD	✓	-	-	31.1	32.34	37.23	33.56	37.79	31.95	39.9	36.55	44.43	42.7	51.17	46.1
URN+MD	✓	-	-	34.6	33.04	39.82	35.82	39.25	36.19	43.78	39.74	45.91	44.02	52.6	47.51
Ours	-	-	-	33.96	25.09	42.09	33.71	33.11	24.61	46.6	34.96	33.51	23.7	52.47	36.56
	✓	-	-	36.26	33.79	41.41	37.15	41.06	36.13	45.53	40.91	47.14	44.86	53.18	48.39
	-	✓	-	32.37	22.92	41.77	32.35	33.99	23.34	<u>48.54</u>	35.29	22.42	33.65	54.32	36.8
	✓	✓	-	35.11	34.69	39.06	36.29	41.78	39.52	46.81	42.7	46.12	**48.22**	53.97	49.44
	✓	✓	✓	**40.05**	**39.93**	**44.1**	**41.36**	**44.62**	**42.7**	**49.05**	**45.46**	**49.28**	46.79	**55.48**	**50.52**

Table 5. Ablation study for our proposed approaches Linear Fusion and M3L. *Our*'s use Linear Fusion as the base segmentation model.

Fusion weight (α)	0.4	0.5	0.6	0.7	0.8	0.9
Val mIoU	56.39	57.30	57.16	56.94	**57.81**	56.30

Table 6. Validation mean IoU on Stanford Indoor dataset (0.2%)

4.3.2 Comparison to Uni-modal Models

To show M3L's uni-modal performance, we train LF with M3L framework utilizing both the modalities during training, however, test with only a single modality at test-time and compare against semi-supervised uni-modal approaches. With RGB only and depth only as inputs, as shown in the Tables 3, our framework shows an absolute improvement of up to 2.96% mIoU for RGB and 5.88% mIoU for depth over CPS-Seg[6] [8], the current state-of-the-art uni-modal segmentation framework. Thus, M3L effectively uses additional modalities for training to even improve the uni-modal segmentation performance and make the models more label efficient.

4.3.3 Linear Fusion

Table 4 shows the performance of our proposed cross-modal fusion method, Linear Fusion, compared to other fusion mechanisms when trained with varying amounts of data. We find that the simpler LF is a strong performer and gives an improvement of up to 3.6% points mIoU over the state-of-the-art TF [42] by using even fewer parameters. Due to the simplicity, LF also has a faster inference time, computed for a data sample on a single Nvidia A40. We also test another set of modalities (RGB+Thermal) using MFNet [11] dataset and show the superiority of our method in appendix Table 10. These results validate the effectiveness of our proposed Linear Fusion for cross-modal fusion.

4.3.4 Ablation

Linear Fusion. As mentioned in Section 3.2, the fusion weight α is a hyperparameter denoting the linear fusion weight for tokens of the two modalities. Hence, we use the validation set and tune α for the Stanford Indoor dataset [1] when trained with 0.2% data. As presented in Table 6, we vary the fusion weight from 0.4 to 0.9 with increments of 0.1 and choose $\alpha = 0.8$. We also see that the performance does not vary much with changes in the fusion weight, reflecting that our framework is not sensitive to α, as long as it is within a reasonable range.

M3L. To show the effectiveness of different components of M3L, we present the performance of different components individually on the Stanford Indoor dataset in Table 5. We first ablate over the simple modality dropout augmentation, which was first proposed in [27] for gesture recognition, and then used in [13, 34] for other domains like audio-video and animated faces. We use the learned token approach presented in Section 3.3 for TF. For URN [18], we follow the modality dropout proposed in [18]. We show that our simple Linear Fusion benefits the most on the MM-Robust metric using just the modality dropout augmentation. Crucially, this augmentation leads to worse performance for the multi-modal input (RGBD) scenario for all three base segmentation models (TF [42], URN [18], and LF). We then ablate over the use of unlabeled data and show that naively using unlabeled data with or without modality dropout aug does not help with the MM-Robust metric over the sup-only baseline. *M3L*, which uses a *M*ulti-modal teacher for *M*asked *M*odality *L*earning, outperforms all other choices.

5. Conclusion

We explore a new problem of semi-supervised multi-modal semantic segmentation and address its two major challenges: limited supervision during training and missing modalities during testing. To tackle these challenges, we propose (a) Linear Fusion, a simple yet effective fusion mechanism that achieves state-of-the-art results with limited supervision, and (b) M3L, a semi-supervised framework that makes the models robust to a realistic scenario of missing modalities and keeps performance better than its uni-modal counterparts even if multiple modalities are not guaranteed at test time. We build a new semi-supervised multi-modal semantic segmentation benchmark and show the effectiveness of our proposed methods against competitive state-of-art.

[6] CPS [8] proposed DeepLabV3+ [7] as base segmentation model, however, we use Segformer [48] for a fair comparison and call it CPS-Seg.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 2, 6, 7, 8, 12

[2] Reza Azad, Nika Khosravi, and Dorit Merhof. Smu-net: Style matching u-net for brain tumor segmentation with missing modalities. In *International Conference on Medical Imaging with Deep Learning*, pages 48–62. PMLR, 2022. 3

[3] Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto. Semi-supervised vision transformers at scale. *arXiv preprint arXiv:2208.05688*, 2022. 1, 6

[4] Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7088–7097, October 2021. 1, 6

[5] Agisilaos Chartsias, Giorgos Papanastasiou, Chengjia Wang, Scott Semple, David E Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. *IEEE transactions on medical imaging*, 40(3):781–792, 2020. 3

[6] Agisilaos Chartsias, Giorgos Papanastasiou, Chengjia Wang, Scott Semple, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. *IEEE Transactions on Medical Imaging*, 40(3):781–792, 2021. 3

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1, 8, 12

[8] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 6, 7, 8, 12, 15

[9] Liuyuan Deng, Ming Yang, Tianyi Li, Yuesheng He, and Chunxiang Wang. Rfbnet: deep multimodal networks with residual fusion blocks for rgb-d semantic segmentation. *arXiv preprint arXiv:1907.00135*, 2019. 2, 3

[10] Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 74–82. Springer, 2019. 2, 3

[11] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017. 8, 12, 15

[12] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019. 2, 3

[13] Ahmed Hussen Abdelaziz, Barry-John Theobald, Paul Dixon, Reinhard Knothe, Nicholas Apostoloff, and Sachin Kajareker. Modality dropout for improved performance-driven talking faces. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 378–386, 2020. 3, 8

[14] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*, 2018. 2, 3

[15] Amod Jog, Aaron Carass, Snehashis Roy, Dzung L Pham, and Jerry L Prince. Random forest regression for magnetic resonance image synthesis. *Medical image analysis*, 35:475–488, 2017. 3

[16] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 429–445. Springer, 2020. 1, 2, 3

[17] Mans Larsson, Erik Stenborg, Lars Hammarstrand, Marc Pollefeys, Torsten Sattler, and Fredrik Kahl. A cross-season correspondence dataset for robust semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[18] Kenneth Lau, Jonas Adler, and Jens Sjölund. A unified representation network for segmentation with missing modalities. *arXiv preprint arXiv:1908.06683*, 2019. 2, 3, 6, 7, 8, 13, 14

[19] Seungyong Lee, Seong-Jin Park, and Ki-Sang Hong. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4990–4999, 2017. 1, 2, 3, 4

[20] Di Lin, Guangyong Chen, Daniel Cohen-Or, Pheng-Ann Heng, and Hui Huang. Cascaded feature network for semantic segmentation of rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 1311–1319, 2017. 1, 4

[21] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 1

[22] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *arXiv preprint arXiv:2203.04838*, 2022. 1, 2, 3

[23] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4258–4267, 2022. 2, 3, 6

[24] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021. 1, 6

[25] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9819–9828, June 2022. 1

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 7

[27] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015. 3, 8

[28] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 1, 2, 3

[29] Andreas Pfeuffer and Klaus Dietmayer. Robust semantic segmentation in adverse weather conditions by means of sensor data fusion. In *2019 22th International Conference on Information Fusion (FUSION)*, pages 1–8, 2019. 3

[30] Anmol Sharma and Ghassan Hamarneh. Missing mri pulse sequence synthesis using multi-modal generative adversarial network. *IEEE transactions on medical imaging*, 39(4):1170–1183, 2019. 3

[31] Yan Shen and Mingchen Gao. Brain tumor segmentation on mri with missing modalities. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 417–428. Springer, 2019. 2, 3

[32] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016. 7, 13

[33] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 6, 12

[34] Rongyi Sun, Borun Chen, Qingyu Zhou, Yinghui Li, Yunbo Cao, and Hai-Tao Zheng. A non-hierarchical attention network with modality dropout for textual response generation in multimodal dialogue systems. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6582–6586, 2022. 3, 8

[35] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019. 2, 3

[36] Monica Sunkara, Srikanth Ronanki, Dhanush Bekal, Sravan Bodapati, and Katrin Kirchhoff. Multimodal semi-supervised learning framework for punctuation prediction in conversational speech. *arXiv preprint arXiv:2008.00702*, 2020. 3

[37] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2, 3, 5, 6, 7, 12, 15

[38] Junjiao Tian, Wesley Cheung, Nathaniel Glaser, Yen-Cheng Liu, and Zsolt Kira. Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5716–5723. IEEE, 2020. 3

[39] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 128(5):1239–1285, 2020. 1, 2, 3, 4, 6

[40] Abhinav Valada, Gabriel Oliveira, Thomas Brox, and Wolfram Burgard. Towards robust semantic segmentation using deep fusion. In *Robotics: Science and systems (RSS 2016) workshop, are the sceptics right*, 2016. 3

[41] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1828–1838, 2020. 3

[42] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 13, 14

[43] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4835–4845. Curran Associates, Inc., 2020. 1, 2, 3, 4, 6

[44] Yikai Wang, Fuchun Sun, Ming Lu, and Anbang Yao. Learning deep multimodal feature representation with asymmetric multi-layer fusion. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3902–3910, 2020. 1, 4

[45] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022. 1, 2, 3, 6

[46] Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. Semi-supervised vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 605–620. Springer, 2022. 1

[47] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2776–2784, 2023. 3

[48] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021. 1, 4, 6, 7, 8

[49] Bo Xiong, Haoqi Fan, Kristen Grauman, and Christoph Feichtenhofer. Multiview pseudo-labeling for semi-supervised learning from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7209–7219, 2021. 3

[50] Biting Yu, Luping Zhou, Lei Wang, Jurgen Fripp, and Pierrick Bourgeat. 3d cgan based cross-modality mr image synthesis for brain tumor segmentation. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 626–630. IEEE, 2018. 3

[51] Yifei Zhang, Olivier Morel, Marc Blanchon, Ralph Seulin, Mojdeh Rastgoo, and Désiré Sidibé. Exploration of deep learning-based multimodal fusion for semantic road scene segmentation. In *VISAPP 2019 14Th international conference on computer vision theory and applications*, 2019. 2, 3

[52] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021. 1, 2, 3

[53] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 1, 2, 3