A. Appendix

A.1. Additional Metrics

We reported the mean IoU metrics in Tables 2 and 3. Here, we also report the other two popular metrics used in segmentation, namely, mean accuracy (mAcc.) and pixel accuracy (Pix. Acc.). Mean accuracy is the average classification accuracy of a class whereas pixel accuracy is macro classification accuracy for all pixels. Table 7 reports the performance for Stanford Indoor [1] dataset. Table 8 reports the performance for SUN RGBD [33]. As seen, on the MM-Robust, which measures the average performance across three testing scenarios, our method outperforms all baselines for all three metrics.

We also report these three metrics for the uni-modal semi-supervised results in Table 9. We can see that even when tested with a single modality, our method performs better than state-of-the-art uni-modal semi-supervised methods on all three metrics. Since CPS [8] was proposed originally with the DeepLabV3+ [7] base segmentation model, we also compare our model with CPS-Dlv3p with ResNet-101 encoder.

A.2. Results on RGB-Thermal

We also test another set of modalities to test the generalizability of the proposed Linear Fusion for semantic segmentation with multiple modalities. Table 10 shows the results of Linear Fusion when comared with uni-modal Segformer and Token Fusion. We use MFNet dataset [11] which contains 1569 images (820 taken at daytime and 749 taken at nighttime). The classification is done on 9 classes (8 objects + background). The results indictate that Linear Fusion outperforms Token Fusion for RGB-Thermal as well hinting towards the generalizability of the approach.

A.3. Qualitative Examples

We also show qualitative results for randomly chosen examples images from the Stanford Indoor [1] dataset. Figure 6 compares different base segmentation multi-modal models with Linear Fusion and the proposed M3L semi-supervised framework with supervised-only and mean teacher [37] frameworks when trained using only 0.1% (49) labels. We can see that when our model is trained with M3L, the segmentation performance is superior to other supervised or semi-supervised baselines.

We also visualize how the segmentation is affected when a modality is missing. In Figures 7, we see that when the missing modality robustness is left untreated (when trained with mean teacher [37]), the performance is sensitive to the presence of both modalities. In the realistic scenario of missing modalities, the performance degrades significantly. However, when the model is trained with our proposed M3L framework, the predictions can hold up the quality even with missing modalities.

In Figures 4, 5, we see an example where the depth modality plays an more important role as the image captures the inside of a room through a door. This information is represented well in the depth modality. If the missing modality problem is left untreated as in the mean teacher [37] framework, when depth is missing during inference, the prediction worsens significantly as seen in Figure 4. However, when treated properly using the proposed M3L framework, even with missing depth, the performance holds up as shown in Figure 5.

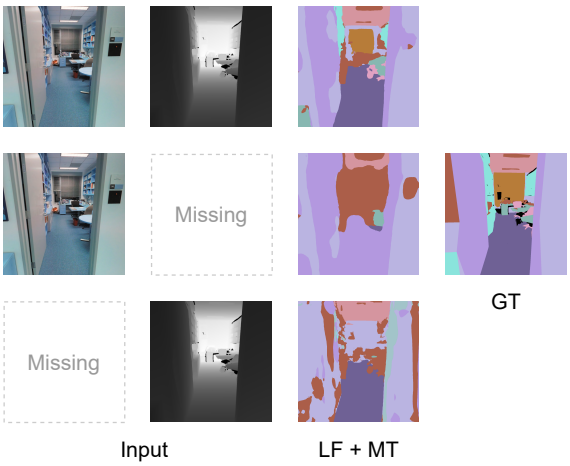

Figure 4. An example to show that when Linear Fusion (LF) is trained with mean teacher (MT) [37], it is sensitive to the presence of both modalities.

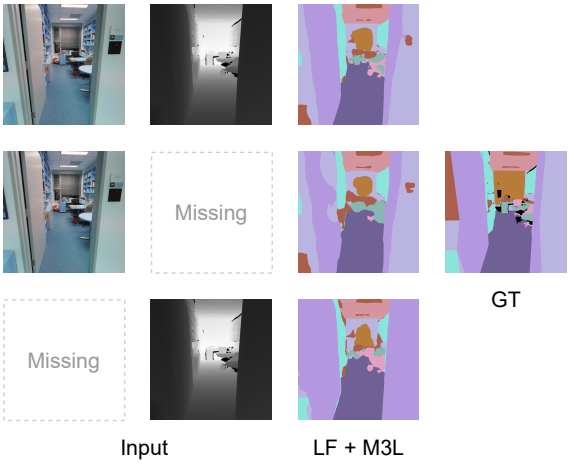

Figure 5. When Linear Fusion is trained with our proposed M3L framework, the predictions are robust to the missing modalities.

A.4. Additional Implementation Details

To train the proposed segmentation model, Linear Fusion with the proposed semi-supervised training framework M3L, we use a batch size of 16 and load 16 labeled and 16 unlabeled data samples in a batch. We calculate the supervised loss on the 16 labeled samples. Since the unsupervised loss is calculated on both the labeled and unlabeled samples and requires a different forward pass on the labeled samples, we make a copy of the labeled samples and compute the unsupervised loss on this copy and the unlabeled samples. Thus, we pass a batch of 48 instances to the model with 16 labeled, 16 labeled (but same examples) and 16 unlabeled, where the masking is done randomly in the last 32 samples of the batch. The ground truth is a single batch of 16 samples (corresponding to the first 16 samples in feed forward). For modality masking in the

Method	RGB			Depth			RGBD			MM-Robust		
	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.
Uni-modal RGB	35.43	47.79	63.93	-	-	-	-	-	-	-		
Uni-modal Depth	-	-	-	34.05	45.63	62.56	-	-	-	-	-	
TF [42]	29.96	42.63	58.41	29.98	41.36	58.82	40.17	50.86	68.82	33.37	44.95	62.02
URN [18]	30.56	44.30	57.82	25.85	37.16	56.07	40.17	52.75	67.23	32.19	44.74	60.37
LF	33.96	47.86	60.06	25.09	36.93	49.93	42.09	**55.25**	69.23	33.71	46.68	59.74
LF + MT	32.37	42.23	59.91	22.92	30.13	56.18	41.77	52.08	68.22	32.35	41.48	61.44
LF + M3L	**40.05**	**50.47**	**69.09**	**39.93**	**49.97**	**70.91**	**44.10**	53.79	**72.94**	**41.36**	**51.41**	**70.98**

(a) 0.1% (49) labeled data

Method	RGB			Depth			RGBD			MM-Robust		
	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.
Uni-modal RGB	39.45	49.97	65.95	-	-	-	-	-	-	-	-	-
Uni-modal Depth	-	-	-	35.24	46.97	64.10	-	-	-	-	-	-
TF [42]	33.11	44.25	60.92	31.47	42.55	59.27	43.04	52.35	70.33	35.87	46.38	63.51
URN [18]	35.71	46.74	62.20	25.14	37.42	56.87	45.87	56.20	70.90	35.57	46.79	63.32
LF	33.51	41.9	60.47	23.7	30.75	54.06	46.6	57.37	71.87	36.56	45	63.74
LF + MT	33.65	42.58	60.92	22.42	29.04	52.71	48.54	57.67	74.85	36.8	45.52	63.77
LF + M3L	**44.62**	**54.99**	**71.28**	**42.70**	**52.60**	**71.91**	**49.05**	**58.28**	**75.01**	**45.46**	**55.29**	**72.73**

(a) 0.2% (98) labeled data

Method	RGB			Depth			RGBD			MM-Robust		
	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.
Uni-modal RGB	46.45	56.2	71.73	-	-	-	-	-	-	-	-	-
Uni-modal Depth	-	-	-	44.78	55.24	72.40	-	-	-	-	-	-
TF [42]	37.34	45.83	65.86	28.33	41.07	57.29	51.85	62.30	75.82	39.17	49.73	66.32
URN [18]	36.25	45.35	64.35	33.27	45.11	62.39	52.07	61.04	76.69	40.53	50.5	67.81
LF	33.51	41.90	60.47	23.70	30.75	54.06	52.47	62.34	76.69	36.56	45.00	63.74
LF + MT	33.65	42.58	60.92	22.42	29.04	52.71	54.32	**64.93**	77.69	36.80	45.52	63.77
LF + M3L	**49.28**	**59.03**	**73.86**	**46.79**	**57.41**	**74.11**	**55.48**	64.78	**78.59**	**50.52**	**60.41**	**75.52**

(a) 1% (491) labeled data

Table 7. We compare the multi-modal models on three testing scenarios: RGBD, RGB (Depth missing), and Depth (RGB missing) using three metrics on Stanford Indoor dataset. We also report the individual uni-modal model's performance for the two modalities for comparison.

student input, we randomly choose either RGB or Depth or None modality to mask. As mentioned, we use the multi-class cross entropy loss for the unsupervised loss and use the OHEM loss [32] as supervised loss with a threshold of 0.7. We ignore the supervised loss for pixels with ground truth class missing. We train our model for 5 epochs (one epoch is defined as passing over all training data, and not just the labeled data, once) for Stanford Indoor dataset and 50 for SUN RGBD dataset which results in 15300 iterations and 14700 iterations respectively for both the datasets, irrespective of the labeled and unlabeled ratio. For training, we scale the images with a random factor between $[0.5, 2]$ and perform a random crop of 500×500 for SUN RGBD and 540×540 for Stanford Indoor. For evaluation, we do a *single-scale*, *non-sliding* evaluation by resizing the original image to the expected model input shape and rescaling the predictions back to the original ground truth shape. The code is implemented using PyTorch's Data Distributed Parallel and was run on 4 Nividia A40 GPUs.

Method	RGB			Depth			RGBD			MM-Robust		
	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.
Uni-modal RGB	28.71	37.21	73.36	-	-	-	-	-	-	-	-	-
Uni-modal Depth	-	-	-	22.81	29.81	70.38	-	-	-	-	-	-
TF [42]	27.97	36.05	72.15	23.58	30.73	70.62	29.31	35.93	74.82	26.95	34.24	72.53
URN [18]	28.72	**39.60**	72.30	12.47	18.00	61.11	31.31	40.54	74.93	24.17	32.71	69.45
LF	29.69	39.17	73.83	15.75	22.24	64.81	**32.00**	**41.48**	75.92	25.81	34.30	71.52
LF + MT	29.57	37.32	74.42	17.86	23.10	67.16	31.11	38.76	76.12	26.18	33.06	72.57
LF + M3L	**29.92**	36.83	**75.19**	**25.44**	**32.30**	**72.32**	30.67	37.20	**76.36**	**28.68**	**35.44**	**74.62**

(a) 6.25% (297) labeled data

Method	RGB			Depth			RGBD			MM-Robust		
	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.
Uni-modal RGB	35.33	45.2	76.15	-	-	-	-	-	-	-	-	-
Uni-modal Depth	-	-	-	27.60	35.54	72.49	-	-	-	-	-	-
TF [42]	33.75	43.99	74.42	28.31	37.04	72.14	35.88	43.93	76.96	32.65	41.65	74.51
URN [18]	33.66	45.67	74.14	15.62	21.68	63.74	37.62	47.55	77.41	28.97	38.30	71.76
LF	35.48	46.32	75.75	17.46	24.29	65.04	39.00	**49.13**	78.20	30.65	39.91	73.00
LF + MT	34.82	45.55	75.57	18.89	28.33	66.75	39.17	47.70	79.02	30.96	40.53	73.78
LF + M3L	**38.12**	**46.93**	**77.80**	**32.29**	**40.96**	**74.91**	**39.70**	47.97	**79.05**	**36.70**	**45.29**	**77.25**

(a) 12.5% (594) labeled data

Method	RGB			Depth			RGBD			MM-Robust		
	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.
Uni-modal RGB	38.31	48.30	77.66	-	-	-	-	-	-	-	-	-
Uni-modal Depth	-	-	-	30.43	38.53	73.70	-	-	-	-	-	-
TF [42]	37.36	48.23	76.15	31.90	40.50	73.79	39.86	48.26	78.67	36.37	45.66	76.20
URN [18]	37.49	49.20	76.24	17.27	22.12	64.68	40.49	50.7	78.87	31.75	40.67	73.26
LF	39.15	50.27	77.53	17.66	25.64	66.67	42.09	**52.32**	79.78	32.97	42.74	74.66
LF + MT	38.96	49.47	77.38	21.03	27.71	68.81	41.95	51.99	79.57	33.98	43.06	75.25
LF + M3L	**41.31**	**51.01**	**79.15**	**34.11**	**42.91**	**75.58**	**42.69**	52.03	**80.4**	**39.37**	**48.65**	**78.38**

(a) 25% (1189) labeled data

Table 8. We compare the multi-modal models on three testing scenarios: RGBD, RGB (Depth missing), and Depth (RGB missing) using three metrics on SUN RGBD dataset. 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)		
	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.
Uni-modal (sup only)	35.43	47.79	63.93	39.45	49.97	65.95	46.45	56.2	71.73
Uni-modal + MT [37]	36.59	46.18	65.43	41.5	52.78	69.16	47.04	56.96	72.6
Uni-modal + CPS-Dlv3p [8]	33.09	42.12	62.56	37.95	48.16	65.8	44.22	53.85	70.81
Uni-modal + CPS-Seg[6] [8]	37.09	48.41	65.97	42.75	51.61	69.96	46.37	56.32	72.86
Ours	**40.05**	**50.47**	**69.09**	**44.62**	**54.99**	**71.28**	**49.28**	**59.03**	**73.86**

(a) RGB uni-modal

Method	0.1 % (49)			0.2% (98)			1% (491)		
	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.
Uni-modal (sup only)	34.05	45.63	62.56	35.24	46.97	64.1	44.78	55.24	72.4
Uni-modal + MT [37]	33.46	43.18	62.45	37.57	48.51	67.5	46.25	56.01	74.68
Uni-modal + CPS-Dlv3p [8]	33.43	43.05	64.97	35.93	47.52	64.11	45.50	55.13	74.04
Uni-modal + CPS-Seg[6] [8]	33.56	42.67	65.75	36.71	47.05	66.64	45.71	55.4	74.32
Ours	**39.93**	**49.97**	**70.91**	**42.7**	**52.6**	**71.91**	**46.79**	**57.41**	**74.11**

(a) Depth uni-modal

Table 9. Uni-modal semi-supervised segmentation. LF when trained with M3L (ours) beats state-of-the-art uni-modal semi-supervised frameworks when tested with a single modality (RGB (a) or Depth (b) modality) as input.

Method	Day			Night		
	mIoU	mAcc.	Pix. Acc.	mIoU	mAcc.	Pix. Acc.
Uni-modal (RGB)	46.32	64.63	97.76	49.73	57.05	97.44
Token Fusion	47.37	63.67	98.04	55.82	63.06	98.11
Linear Fusion (ours)	**48.51**	**66.65**	**98.09**	**57.72**	**66.11**	**98.17**

Table 10. Linear Fusion results with RGB-Thermal modalities on MFNet dataset [11].

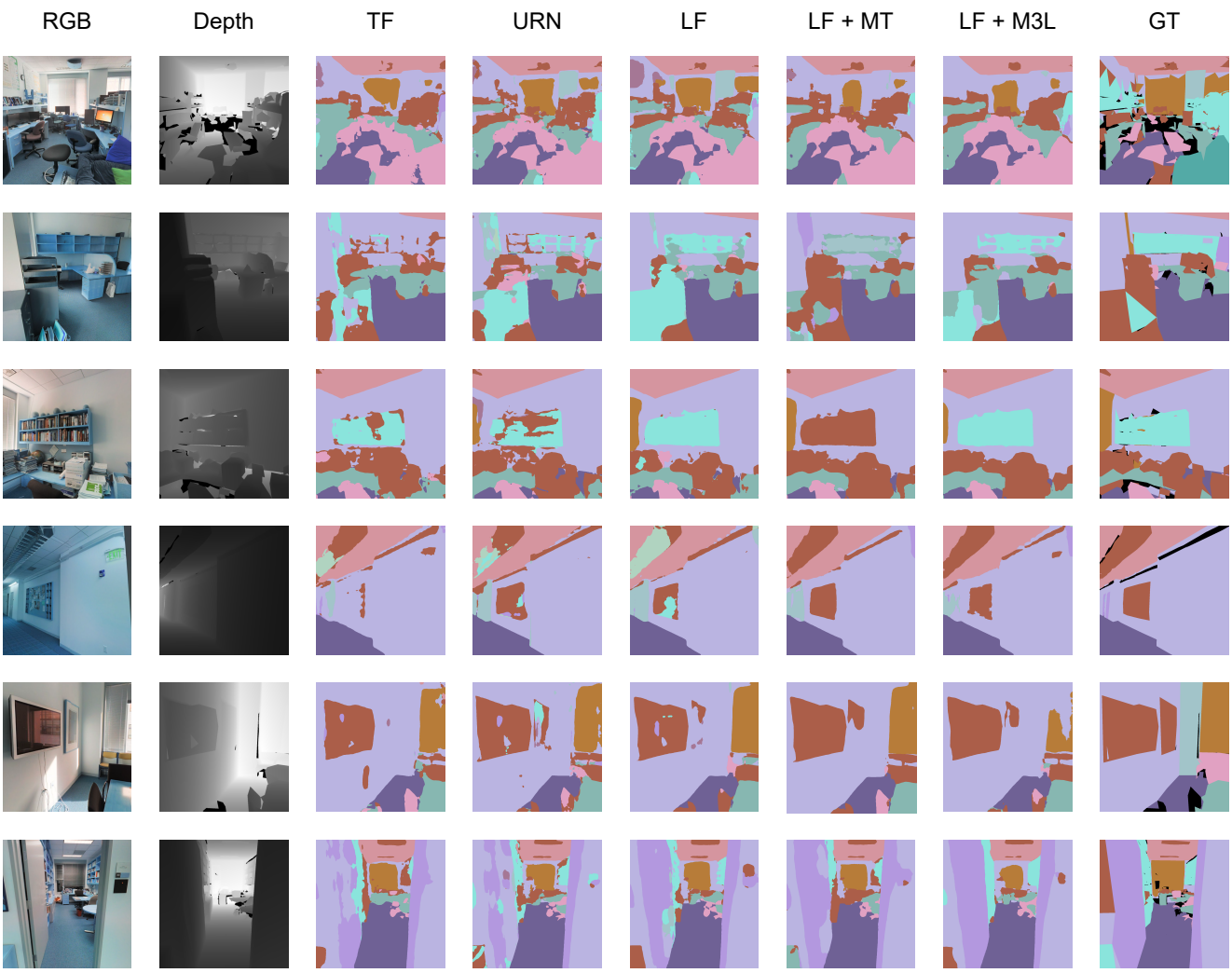

Figure 6. Examples for different multi-modal models trained with supervised and semi-supervised frameworks.

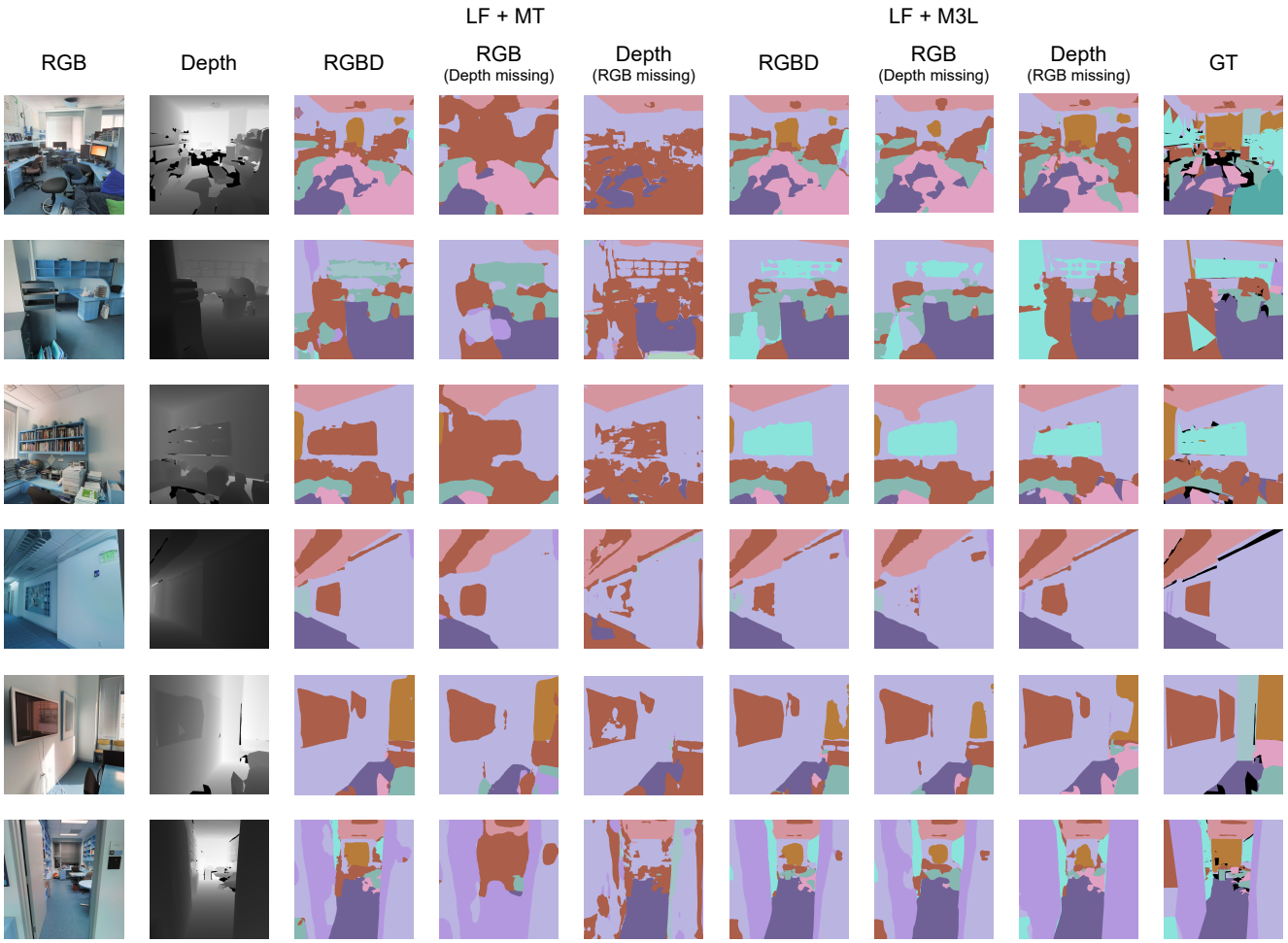

Figure 7. Examples for visualizing drop in performance when a modality is missing and robustness to missing modality when trained with the propsed M3L framework.