Supplementary Material: Understanding Dark Scenes by Contrasting Multi-Modal Observations

Xiaoyu Dong[1,2] and Naoto Yokoya[1,2,✉]
[1]The University of Tokyo, Japan
[2]RIKEN AIP, Japan
dong@ms.k.u-tokyo.ac.jp, yokoya@k.u-tokyo.ac.jp

Section I provides more ablation studies. Section II provides more qualitative results. Section III provides additional discussions and experiments. Experiments in the supplementary material are conducted using our model with the SegNeXt-B backbone.

I. More Ablation Studies

I.I. Ablations on Hyperparameters

Our supervised multi-modal contrastive learning approach involves four hyperparameters, including loss weights λ_{cm}, λ_{vis}, λ_{aux} (in Eq. (7)), and temperature τ (see Eqs. (4) and (6)).

$\lambda_{cm}, \lambda_{vis}, \lambda_{aux}$	0	0.01	0.05	0.1	0.2
	mIoU (%)				
Low	66.02	68.01	67.54	68.36	**68.76**
Night	58.44	59.48	**60.00**	58.78	58.60
Normal	54.70	55.70	**55.77**	55.02	55.32

Table I. Ablations on loss weights λ_{cm}, λ_{vis}, and λ_{aux} for low-light indoor, nighttime outdoor, and normal-light scene segmentation. We set $\lambda_{cm} = \lambda_{vis} = \lambda_{aux}$ to balance the effect of our three contrastive losses. Single-scale results are reported. The best results are shown in **bold**.

τ	0.05	0.1	0.2
	mIoU (%)		
Low	68.19	**68.76**	68.37
Night	59.67	**60.00**	58.50
Normal	55.43	**55.77**	55.46

Table II. Ablations on temperature τ for low-light indoor, nighttime outdoor, and normal-light scene segmentation tasks. Single-scale results are reported. The best results are shown in **bold**.

Table I reports the ablation studies on λ_{cm}, λ_{vis}, and λ_{aux}. τ is set as 0.1 in the experiments. As shown, a proper setting of loss weights enhances the effect of our approach. Therefore, according to the results, we set λ_{cm}, λ_{vis}, and λ_{aux} as 0.2 and 0.05 in our experiments for low-light indoor scene segmentation and experiments for nighttime outdoor and normal-light scene segmentation, respectively.

Based on the selected loss weights, we study the effect of τ in Tab. II. According to the ablation results, which show that a middle temperature setting allows better performance, we set τ as 0.1 in our three tasks.

I.II. Ablations on Intermediate Modules

Our intermediate modules follow common protocols to learn spatial and channel weights (coefficient matrix and vector) from multi-modal features to model their dependency and facilitate their interaction. However, different from previous methods [1, 7], which learn separate weights for different modalities, we learn a shared weight (see Eqs. (1) and (2)) to better model the correspondence of two modalities both carrying limited cues. Here we first conduct basic ablations to study the effect of our intermediate modules, then compare the different weight learning manners to justify our superiority.

	Model_a	Model_b	Model_c	Model_d
Spatial Learning	✗	✓	✗	✓
Channel Learning	✗	✗	✓	✓
mIoU (%)	64.49	65.15	65.45	**66.02**

Table III. Effect of our intermediate modules. Single-scale results on low-light indoor scenes are reported. The best result is shown in **bold**. Model_d is identical to the Model_1 in Tab. 5.

Basic Ablations. We study our intermediate modules in Tab. III. Our supervised multi-modal contrastive learning approach is not used here. First, we remove our four modules to build Model_a, in which there is no interaction with features from the two encoders before feeding to the decoder. The segmentation accuracy is only 64.49%. Then

we add four modules with only spatial or channel coefficient learning to build Model_b or Model_c, in which features from the two encoders have spatial or channel interaction (see Eqs. (1) and (2)). The accuracy is 65.15% or 65.45%, respectively. After combining two kinds of learning in Model_d to allow both spatial and channel interaction between multi-modal features as in Eqs. (1) and (2), the accuracy rises to 66.02%.

Manner	Low	Night	Normal
	mIoU (%)		
Separate	68.27	59.67	55.30
Shared	**68.76**	**60.00**	**55.77**

Table IV. Comparison of weight learning manners. Our full model is employed. Single-scale results are reported. The best results are shown in **bold**.

Comparison of Learning Manners. Different from the interaction modules of previous methods [1, 7], which learn separate weights for for different modalities, our intermediate modules learn a shared weight to better model the correspondence of the limited cues in the visible and auxiliary modalities. This follows the idea of our supervised multi-modal contrastive learning approach, which aims to fully capture the correspondence between cross-modal contextual and geometry cues. Based on our full model trained by adding the proposed contrastive approach, we compare our weight learning manner (denoted as Shared) and the learning manner of previous methods (denoted as Separate) in Tab. IV. Experiments on three tasks consistently demonstrate our effectiveness and superiority.

I.III. Efficiency Evaluation

Table V reports the model efficiency. By comparing the first three models, we can observe that the spatial/channel coefficient learning in the four intermediate modules uses $3.09/13.87$M parameters and $2.87/0.02$G FLOPs in total. By comparing the last two models, we can see that our supervised multi-modal contrastive learning approach can increase model performance while maintains the efficiency.

Model	Params. (M)	FLOPs (G)	Low
			mIoU (%)
Only Spa.	57.68	72.21	65.15
Only Cha.	68.46	69.36	65.45
Spa. + Cha.	71.55	72.23	66.02
S + C + SMMCL	71.55	72.23	**68.76**

Table V. Evaluation of model efficiency on low-light indoor scenes. Input visible and auxiliary images are of size $3\times480\times640$. Single-scale results are reported. The best result is shown in **bold**.

II. More Qualitative Results

Figure I presents more qualitative comparisons between our model, the baseline, and three state-of-the-art multi-modal image segmentation methods (CMX [7], TokenFusion [5], and CEN [6]) on the task of understanding low-light indoor scenes from RGB-depth data. Figure II visualizes more results from our model, the baseline, TokenFusion [5], and two RGB-thermal segmentation methods (CMX [7] and ABMDRNet [8]) for the nighttime outdoor scene segmentation task. As can be seen in the comparisons between our model and the baseline, our supervised multi-modal contrastive learning approach can effectively enhance dark scene understanding based on multi-modal images with limited semantics. Besides, compared with those state-of-the-art methods, which develop advanced fusion techniques but neglect the importance of class correlations in multi-modal dark scene semantic segmentation, our model can predict more accurate class information for objects in darkness, such as the pillow in the second scene of Fig. I and the bicycles in the second scene of Fig. II, and achieves much higher overall accuracy in the tasks.

Figure III further provides more qualitative comparisons between our model, the baseline, CMX [7], TokenFusion (SegFomer-B3) [5], and CEN [6] for the RGB-depth based normal-light scene segmentation task. As can be observed, our supervised multi-modal contrastive learning approach shows superior generalization capability in normal-light scenarios, and enables our model to achieve segmentation masks closer to the ground truth and create a state of the art in multi-modal scene understanding.

III. Additional Discussions and Experiments

III.I. Robustness to "Invalid" Auxiliary Modality

An interesting phenomenon in multi-modal scene understanding is that the auxiliary modality might not provide effective geometry cues for objects. This would cause a learning bias problem of segmentation models.

We show example scenarios in the task of dark scene understanding in Fig. IV. In the first scene, the depth values of different objects are very similar. In the second scene, the thermal response of the car is similar to that of the background objects. Segmentation models tend to learn a bias towards the auxiliary modality and confuse those objects in the first scene and the car and background in the second scene (we show results from two state-of-the-art methods and the baseline). This problem exists in even normal-light scene understanding. As shown in Fig. V, those models tend to confuse the bicycles and background in the first scene and the fireplace and wall in the second scene. In contrast, our model can better avoid this learning bias problem by enjoying the contextual cues in the visible modality, showing stronger robustness to "invalid" auxiliary modal-

ity and higher discriminative learning capability to different image modalities. This is thanks to our supervised multi-modal contrastive learning approach, which fully considers the cross-modal context-geometry correspondence and effectively boosts the learning on the visible and auxiliary modalities in our tasks.

III.II. Failure Case

The adoption of our supervised multi-modal contrastive learning approach enables our model to achieve higher accuracy when understanding dark scenes based on multi-modal images with limited semantic information and show stronger robustness to auxiliary modalities that do not provide effective geometry cues. However, in our tasks, we see failure case when both the visible and auxiliary modalities do not provide effective spatial cues for objects. Examples are shown in Fig. VI. Our model fails at the bicycles between the two persons in the first scene and a black car in the second scene, because the RGB modality do not provide effective contextual cues and the thermal modality do not provide effective geometry cues. This happens to other methods as well. Another example is the bicycle in the third scene of Fig. II. Advanced image synthesis techniques [2] might be helpful for such challenging case.

III.III. Application on Other Segmentation Models

We apply our supervised multi-modal contrastive learning approach on SA-Gate [1] and CMX [7] and report the quantitative results in Tab. VI. The loss weights λ_{cm}, λ_{vis}, and λ_{aux} are set as 0.05 in the three tasks. As shown, with the addition of our approach, the accuracy achieved by SA-Gate and the accuracy achieved by CMX in low-light indoor/nighttime outdoor/normal-light scene segmentation increase to $63.18\%/57.79\%/52.30\%$ and $67.58\%/59.63\%/55.28\%$, respectively.

Method	Backbone	Low	Night	Normal
		mIoU (%)		
SA-Gate	ResNet-101	61.79	56.35	51.45
SA-G + SMMCL	ResNet-101	63.18	57.79	52.30
CMX	SegFormer-B2	66.52	57.80	54.10
CMX + SMMCL	SegFormer-B2	**67.58**	**59.63**	**55.28**

Table VI. Application of our supervised multi-contrastive learning approach on other segmentation models. Single-scale results are reported. The best results are shown in **bold**.

Qualitative results from CMX and CMX + SMMCL are further provided in Fig. VII. As can be seen, by applying our approach to include a consideration for class correlations during loss optimization, CMX can achieve more accurate class predictions in both dark and normal-light scenes based on different image modalities. This demonstrates again the effectiveness, generalizability, and applicability of our approach in dark scene understanding and multi-modal scene understanding tasks.

References

[1] 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. In *ECCV*, 2020. 1, 2, 3

[2] Chaowei Fang, Liang Wang, Dingwen Zhang, Jun Xu, Yixuan Yuan, and Junwei Han. Incremental cross-view mutual distillation for self-supervised medical CT synthesis. In *CVPR*, 2022. 3

[3] Saurabh Gupta, Ross Girshick, Pablo Arbelaez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV*, 2014. 6

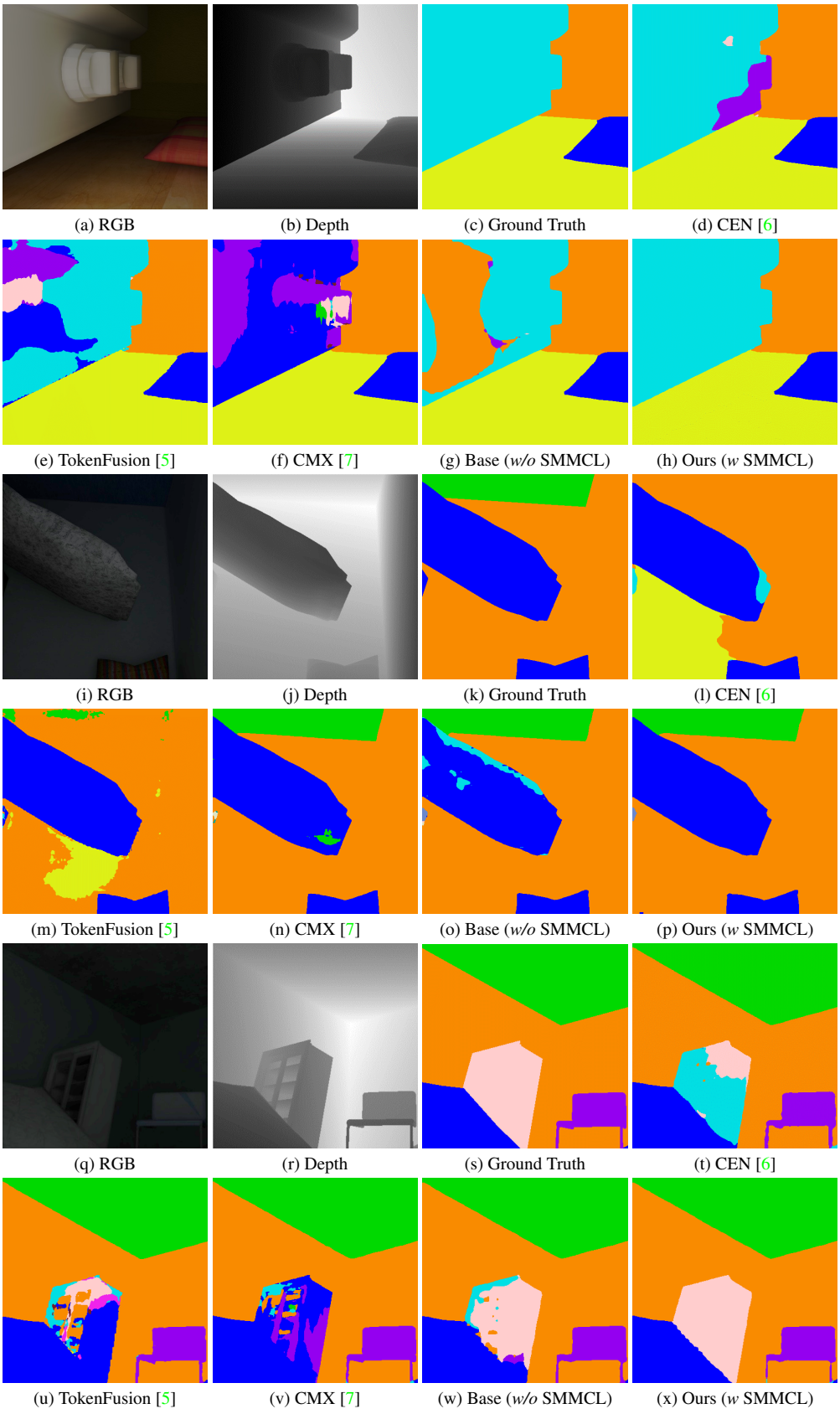
[4] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IROS*, 2017. 7

[5] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *CVPR*, 2022. 2, 4, 5, 6, 7, 8

[6] Yikai Wang, Fuchun Sun, Wenbing Huang, Fengxiang He, and Dacheng Tao. Channel exchanging networks for multi-modal and multitask dense image prediction. *IEEE TPAMI*, 2022. 2, 4, 6

[7] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[8] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. ABMDRNet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation. In *CVPR*, 2021. 2, 5



(a) RGB (b) Depth (c) Ground Truth (d) CEN [6]

(e) TokenFusion [5] (f) CMX [7] (g) Base (*w/o* SMMCL) (h) Ours (*w* SMMCL)

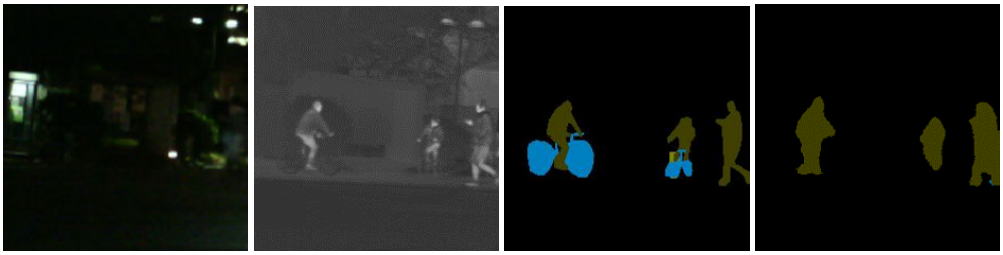
(i) RGB (j) Depth (k) Ground Truth (l) CEN [6]

(m) TokenFusion [5] (n) CMX [7] (o) Base (*w/o* SMMCL) (p) Ours (*w* SMMCL)

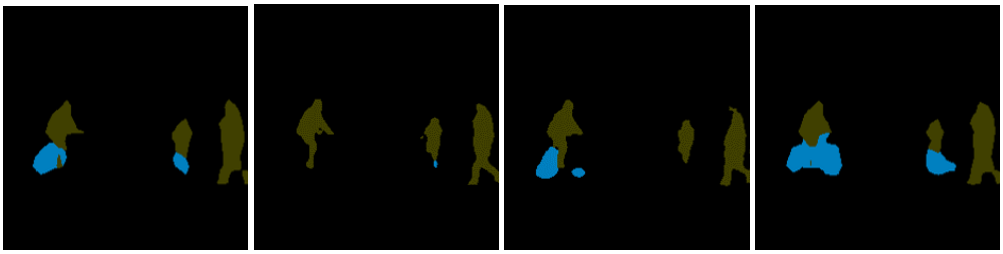
(q) RGB (r) Depth (s) Ground Truth (t) CEN [6]

(u) TokenFusion [5] (v) CMX [7] (w) Base (*w/o* SMMCL) (x) Ours (*w* SMMCL)

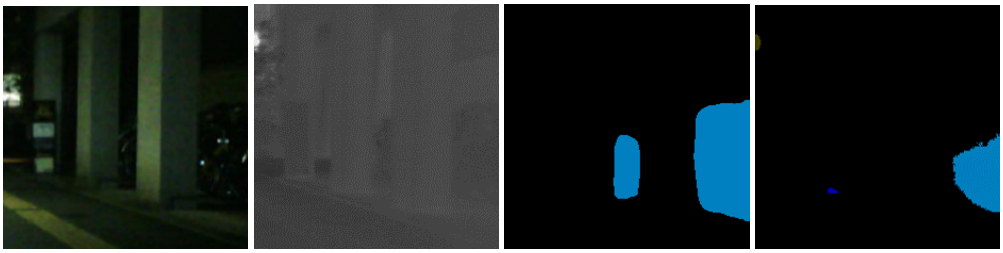
Figure I. Low-light indoor scene segmentation from RGB-depth data.



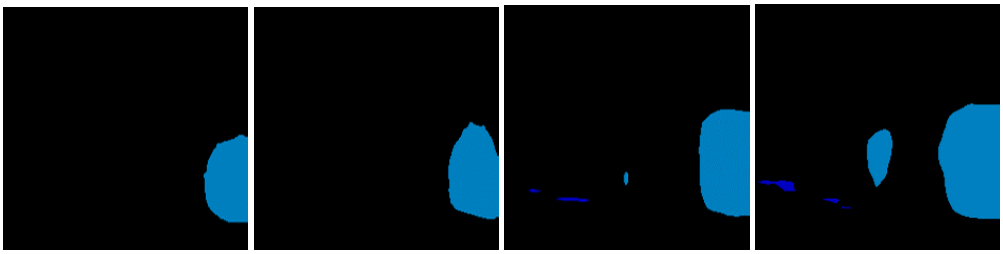
(a) RGB (b) Thermal (c) Ground Truth (d) ABMDRNet [8]



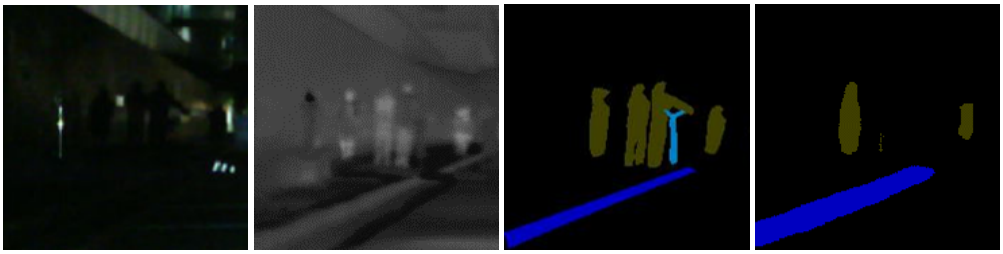
(e) TokenFusion [5] (f) CMX [7] (g) Base (*w/o* SMMCL) (h) Ours (*w* SMMCL)



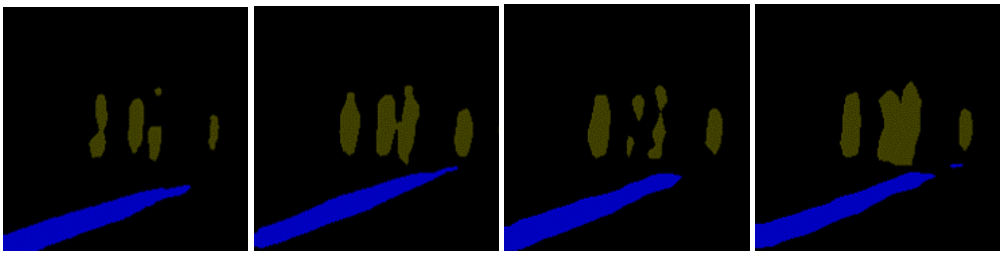
(i) RGB (j) Thermal (k) Ground Truth (l) ABMDRNet [8]



(m) TokenFusion [5] (n) CMX [7] (o) Base (*w/o* SMMCL) (p) Ours (*w* SMMCL)

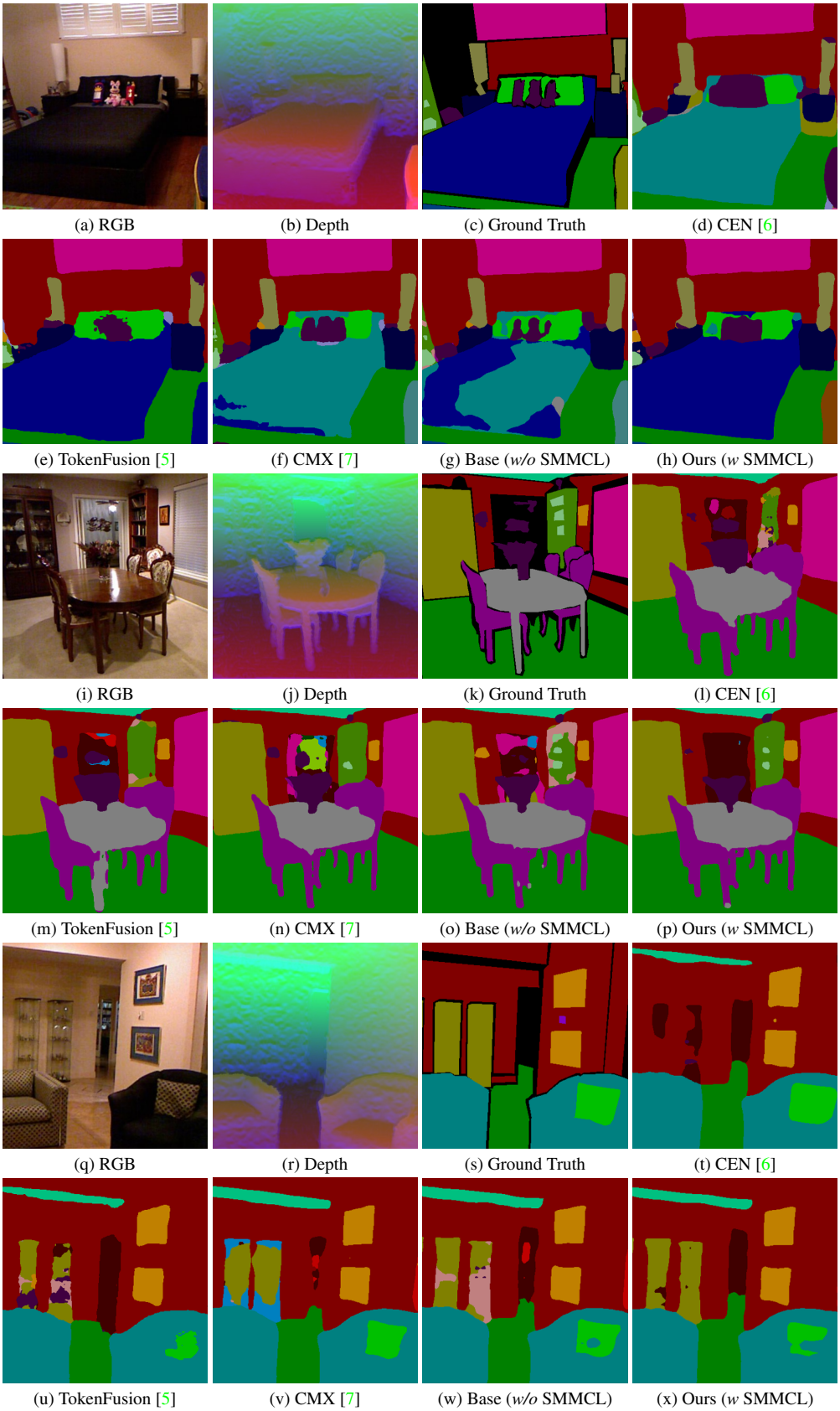


(q) RGB (r) Thermal (s) Ground Truth (t) ABMDRNet [8]



(u) TokenFusion [5] (v) CMX [7] (w) Base (*w/o* SMMCL) (x) Ours (*w* SMMCL)

Figure II. Nighttime outdoor scene segmentation from RGB-thermal data.



(a) RGB (b) Depth (c) Ground Truth (d) CEN [6]

(e) TokenFusion [5] (f) CMX [7] (g) Base (*w/o* SMMCL) (h) Ours (*w* SMMCL)

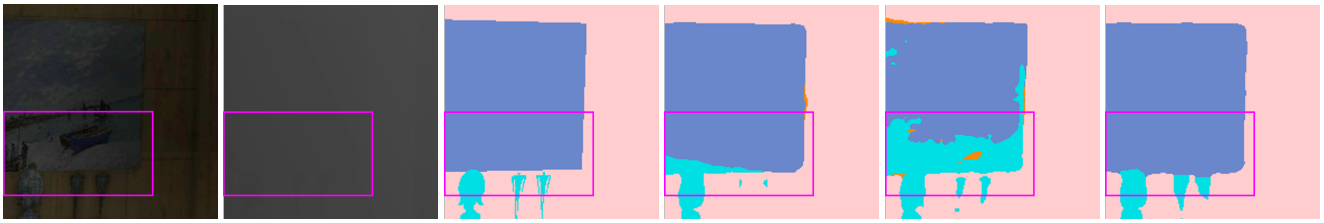
(i) RGB (j) Depth (k) Ground Truth (l) CEN [6]

(m) TokenFusion [5] (n) CMX [7] (o) Base (*w/o* SMMCL) (p) Ours (*w* SMMCL)

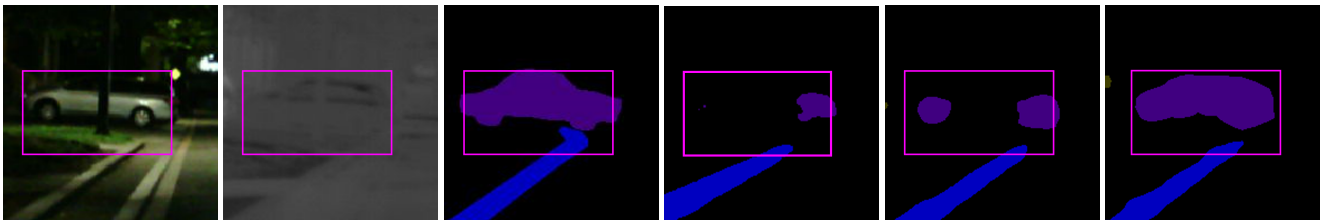
(q) RGB (r) Depth (s) Ground Truth (t) CEN [6]

(u) TokenFusion [5] (v) CMX [7] (w) Base (*w/o* SMMCL) (x) Ours (*w* SMMCL)

Figure III. Normal-light scene segmentation from RGB-depth data. Depth images are encoded to HHA maps [3] in this task.

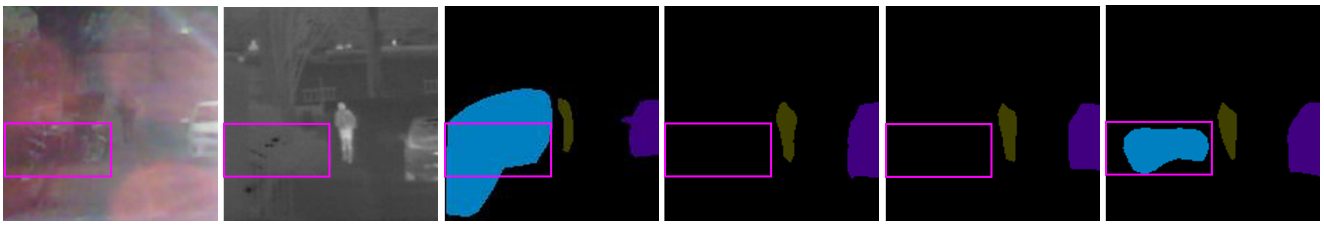


(a) RGB (b) Depth (c) Ground Truth (d) CMX [7] (e) Base (*w/o* SMMCL) (f) Ours (*w* SMMCL)

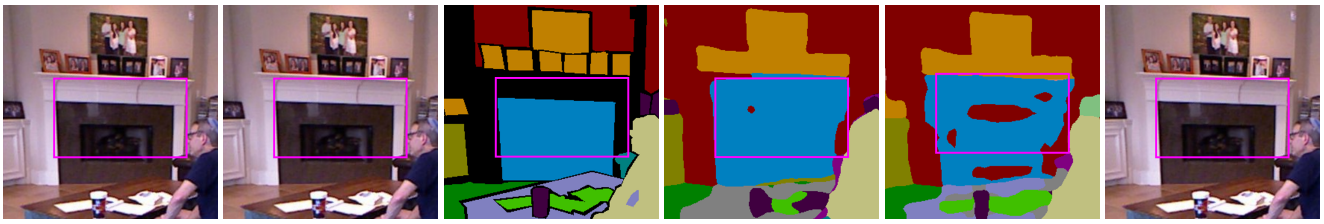


(g) RGB (h) Thermal (i) Ground Truth (j) TokenFusion [5] (k) Base (*w/o* SMMCL) (l) Ours (*w* SMMCL)

Figure IV. Robustness to "invalid" auxiliary modality in dark scene understanding.

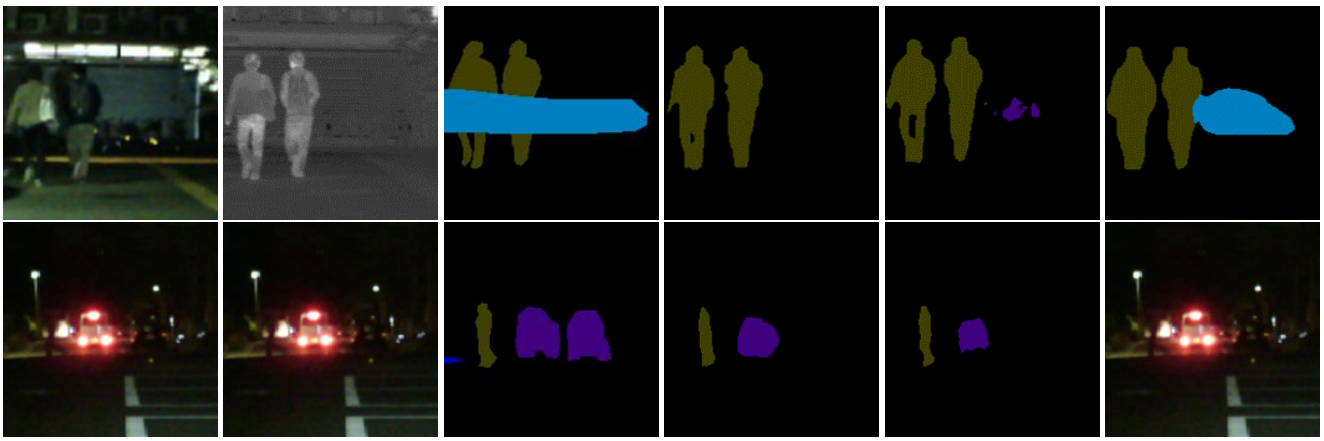


(a) RGB (b) Thermal (c) Ground Truth (d) CMX [7] (e) Base (*w/o* SMMCL) (f) Ours (*w* SMMCL)



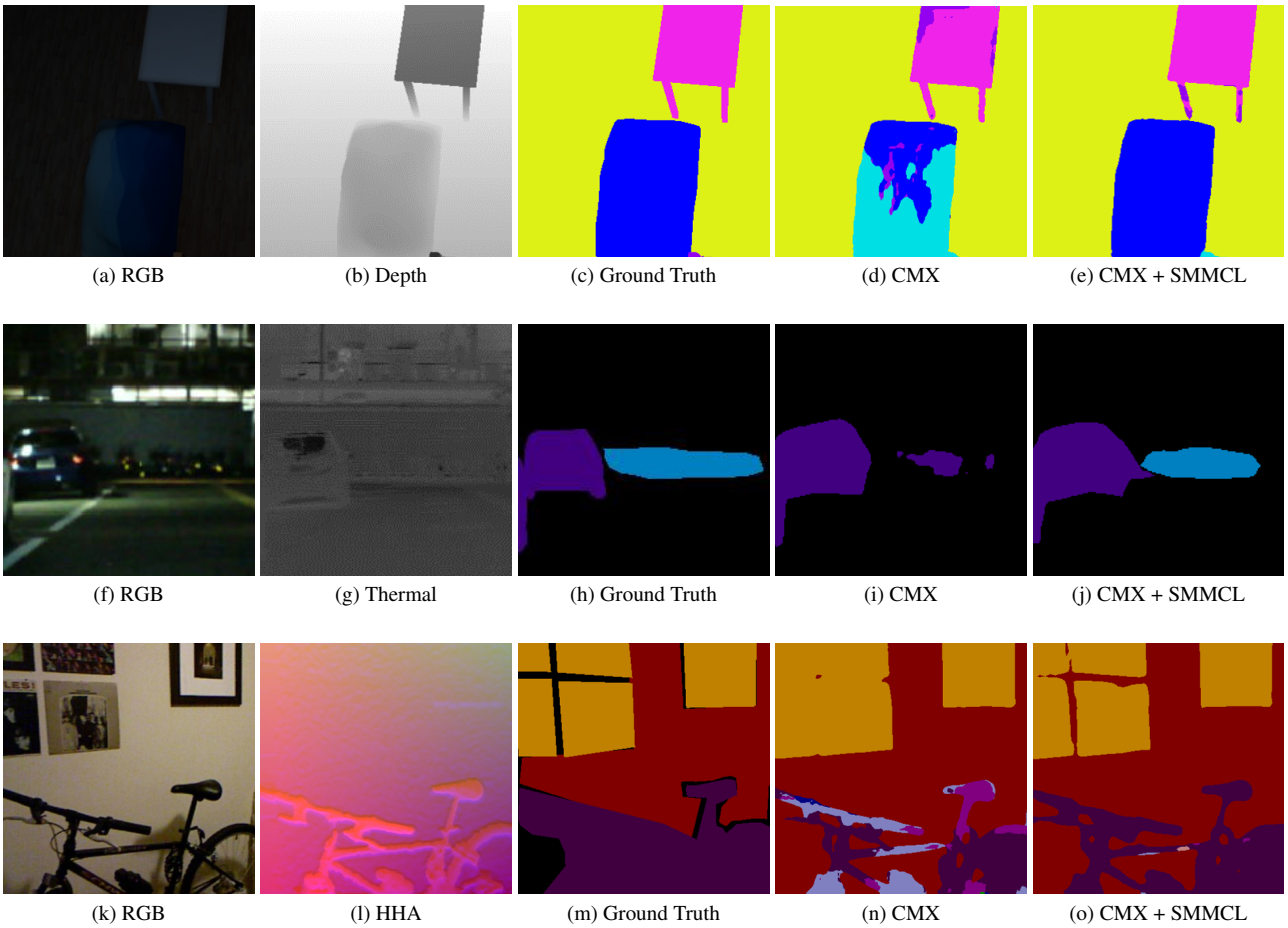
(g) RGB (h) Depth (i) Ground Truth (j) TokenFusion [5] (k) Base (*w/o* SMMCL) (l) Ours (*w* SMMCL)

Figure V. Robustness to "invalid" auxiliary modality in normal-light scene understanding. The first row shows a scene from the daytime split in the test set of the MFNet dataset [4].



(a) RGB (b) Depth (c) Ground Truth (d) TokenFusion [5] (e) CMX [7] (f) Ours

Figure VI. Failure case in nighttime scene segmentation from RGB-thermal data. Note that, the ground truth of the first scene is not exactly correct: there should be bicycles between the two persons; no bicycles on their bodies.



(a) RGB (b) Depth (c) Ground Truth (d) CMX (e) CMX + SMMCL

(f) RGB (g) Thermal (h) Ground Truth (i) CMX (j) CMX + SMMCL

(k) RGB (l) HHA (m) Ground Truth (n) CMX (o) CMX + SMMCL

Figure VII. Application of our supervised multi-modal contrastive learning approach on CMX [7].