

TUNI: Real-time RGB-T Semantic Segmentation with Unified Multi-Modal Feature Extraction and Cross-Modal Feature Fusion

Xiaodong Guo, Tong Liu, Yike Li, Zi'ang Lin, Zhihong Deng

Abstract—RGB-thermal (RGB-T) semantic segmentation improves the environmental perception of autonomous platforms in challenging conditions. Prevailing models employ encoders pre-trained on RGB images to extract features from both RGB and infrared inputs, and design additional modules to achieve cross-modal feature fusion. This results in limited thermal feature extraction and suboptimal cross-modal fusion, while the redundant encoders further compromises the model’s real-time efficiency. To address the above issues, we propose TUNI, with an RGB-T encoder consisting of multiple stacked blocks that simultaneously perform multi-modal feature extraction and cross-modal fusion. By leveraging large-scale pre-training with RGB and pseudo-thermal data, the RGB-T encoder learns to integrate feature extraction and fusion in a unified manner. By slimming down the thermal branch, the encoder achieves a more compact architecture. Moreover, we introduce an RGB-T local module to strengthen the encoder’s capacity for cross-modal local feature fusion. The RGB-T local module employs adaptive cosine similarity to selectively emphasize salient consistent and distinct local features across RGB-T modalities. Experimental results show that TUNI achieves competitive performance with state-of-the-art models on FMB, PST900 and CART, with fewer parameters and lower computational cost. Meanwhile, it achieves an inference speed of 27 FPS on a Jetson Orin NX, demonstrating its real-time capability in deployment. Codes are available at <https://github.com/xiaodonguo/TUNI>.

I. INTRODUCTION

The intelligence of autonomous platforms (e.g., mobile robots, drones, autonomous cars) relies heavily on the perception and understanding of their surroundings. RGB-thermal semantic segmentation performs pixel-level classification on images, utilizing Red-Green-Blue (RGB) and thermal information that captured by visible light and long-wave infrared cameras, respectively. RGB information provides rich color and texture cues, while thermal information captures heat distribution, serving as a complement in challenging scenarios, such as low-illumination, glare and low color contrast [1], [2].

Unlike single-modal segmentation models, RGB-T segmentation models need to extract information from two modalities while realize cross-modal feature fusion. As shown in Fig. 1 (a), prevailing methods [3]–[5] typically follow a pipeline consisting of independent feature extraction, cross-modal feature fusion, and decoding. Specifically, modality-specific features are extracted from RGB and thermal images using the backbone pre-trained on ImageNet [6],

*This work was supported by National Natural Science Foundation of China under Grant 62476026. (Corresponding author: Tong Liu).

Xiaodong Guo, Tong Liu, Yike Li, Zi'ang Lin and Zhihong Deng are with the School of Automation, Beijing Institute of Technology, Beijing 100081, China (e-mail: liutong2002@bit.edu.cn, 3120245534@bit.edu.cn).

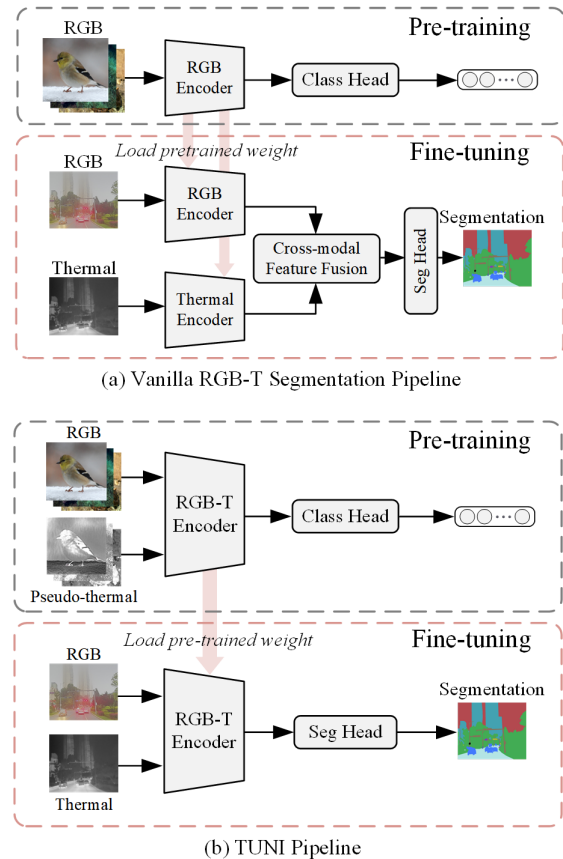


Fig. 1. Two RGB-T semantic segmentation pipelines: (a) vanilla RGB-T segmentation pipeline, (b) TUNI pipeline. The TUNI encoder unifies multi-modal feature extraction and cross-modal fusion within a more compact architecture.

followed by dedicated cross-modal feature fusion modules deployed on four encoder stages. Although effective, this pipeline suffers from three primary limitations, including 1) encoders pre-trained on RGB images are suboptimal for thermal feature extraction, 2) cross-modal feature fusion only on four encoder stages are insufficient, and 3) duplicated feature extraction structures introduce considerable redundancy, substantially increasing the overall model complexity.

The first limitation stems from the modality constraint in the pre-training phase. Large-scale datasets are fundamental for training effective feature extractors; however, the absence of thermal data makes it infeasible to the training on the thermal modality. Alternatively, RGB pre-trained feature extractors are utilized to extract thermal features, which is suboptimal due to the heterogeneous nature of RGB

and thermal images. The second limitation arises from the independence between the encoders and cross-modal feature fusion modules. Commonly adopted strategies are to perform cross-modal feature fusion at the four hierarchical feature levels produced by encoders. However, a deep network contains many feature extraction blocks between successive layers, and limiting cross-modal fusion to only four stages is insufficient to fully exploit the complementary information across modalities. The final limitation is attributed to the inherent information asymmetry between RGB and thermal data. As thermal images convey less information than RGB images, employing the same encoder is redundant. It is imperative to eliminate redundant components for efficient deployment on edge devices with limited resources.

Some of these limitations have been partially resolved by prior studies; however, their solutions remain incomplete and leave room for further improvement. Dong et al. [7] replace the thermal encoders with multimodal prompt generation modules, therefore reduce the complexity of the thermal branch. Nevertheless, these modules are only applied to four RGB encoder layers, which limits the fusion effectiveness. Liu et al. [8] propose a multi-modal cross-prompt module and insert it between each pair of RGB and thermal encoder blocks, while the duplicated thermal encoder introduces redundancy. Zhou et al. [9], [10] adopt asymmetric lightweight encoders to extract thermal features, such as Mit-B0 [11] and MobileNet [12]. Although the model complexity is reduced, the lightweight encoders pre-trained on RGB images lead to limited capability to extract thermal feature. Moreover, current research often achieves high model performance at the cost of increased parameters and inference burden, while overlooking deployment issues that are crucial for real-world robotic applications.

To comprehensively address the aforementioned challenges, we propose the TUNI for real-time RGB-T semantic segmentation in robotic applications. As shown in Fig 1 (b), we propose an RGB-T encoder, which is used in both pre-training and fine-tuning phases. This encoder is composed of multiple stacked blocks, each of which simultaneously performs multi-modal feature extraction and cross-modal feature fusion. In the pre-training phase, an RGB-T image translation model [13] is utilized to generate pseudo-thermal images. Using RGB and pseudo-thermal data, the RGB-T encoder undergoes large-scale pre-training to jointly learn multi-modal feature extraction and cross-modal feature fusion. Furthermore, by slimming down the thermal branch, the encoder achieves a more compact architecture.

This idea is primarily motivated by DFormer [14], which introduces RGB-depth (RGB-D) pre-training for RGB-D semantic segmentation. Except for the different data preparation for the thermal modality, we further develop the network architecture. Specifically, each encoder block consists of three modules: RGB-RGB local module, RGB-T global module and RGB-T local module. The first two modules are inherited from DFormer and are respectively used for RGB local feature extraction and RGB-T global feature fusion. In contrast with the Hamilton product for coarse local cross-

modal feature fusion used in DFormer, we further consider the attribute differences between RGB-T information. The RGB-T local module employs adaptive cosine similarity to selectively emphasize salient consistent and distinct local features across RGB-T modalities, achieving precise cross-modal local feature fusion.

Our main contributions can be summarized as follows:

- We propose a real-time RGB-T semantic segmentation model, TUNI, for robotic application. With RGB-T pre-training, TUNI performs multi-modal feature extraction while enhancing cross-modal feature fusion.
- We propose an RGB-T local module for precise cross-modal local feature fusion. It employs adaptive cosine similarity to selectively emphasize salient consistent and distinct local features across RGB-T modalities.
- We evaluate the performance of TUNI on FMB, PST900, and CART datasets, demonstrating its competitive performance compared to state-of-the-art (SOTA) methods with a more compact architecture. Moreover, TUNI achieves an inference speed of 120 frames per second (FPS) on an RTX 4090 and 27 FPS on a Jetson Orin NX.

II. RELATED WORK

The goal of RGB-T semantic segmentation is to exploit the synergy between RGB and thermal modalities, thus improving the robustness of the segmentation in challenging scenarios. In current RGB-T semantic segmentation research, encoders pre-trained on RGB modality (e.g., ResNet [15], MiT [11], and Swin-Transformer [16]) are commonly employed to extract features from both RGB and thermal inputs. Subsequently, dedicated cross-modal fusion modules are designed to integrate RGB and thermal features in fine-tune phase. Zhou et al. [3] employ spatial attention on low-level features and channel attention on high-level features to achieve adaptive cross-modal feature fusion. Zhang et al. [5], [17] introduce cross-attention to construct long-range dependency between RGB-T features. Guo et al. [4] propose a multi-attribute hierarchical feature interaction module to realize multi-scale cross-modal feature fusion. Liu et al. [18] propose a multiplex interactive learning model, which integrates multi-model, multi-modal and multi-level feature learning. Zhou et al. [19] combine multi-modal features via gating mechanisms and enhance spatial details via introducing edge information. In addition, prompt learning has been introduced to address information asymmetry and reduce encoder complexity. Dong et al. [7] present two prompt learning modules to fuse cross-modal feature and adapt prompted multi-modal features in the frozen encoder. Liu et al. [8] develop a framework that exploits the complementarity and bridges the semantic gap between modalities. To further enhance the encoder’s ability to extract features from auxiliary modalities, Yin et al. [14] propose DFormer, which generates pseudo-depth images during the pre-training stage to train an RGB-depth encoder. Nevertheless, its effectiveness in RGB-T tasks remains unverified, and the fusion of cross-modal local features in DFormer remains relatively coarse.

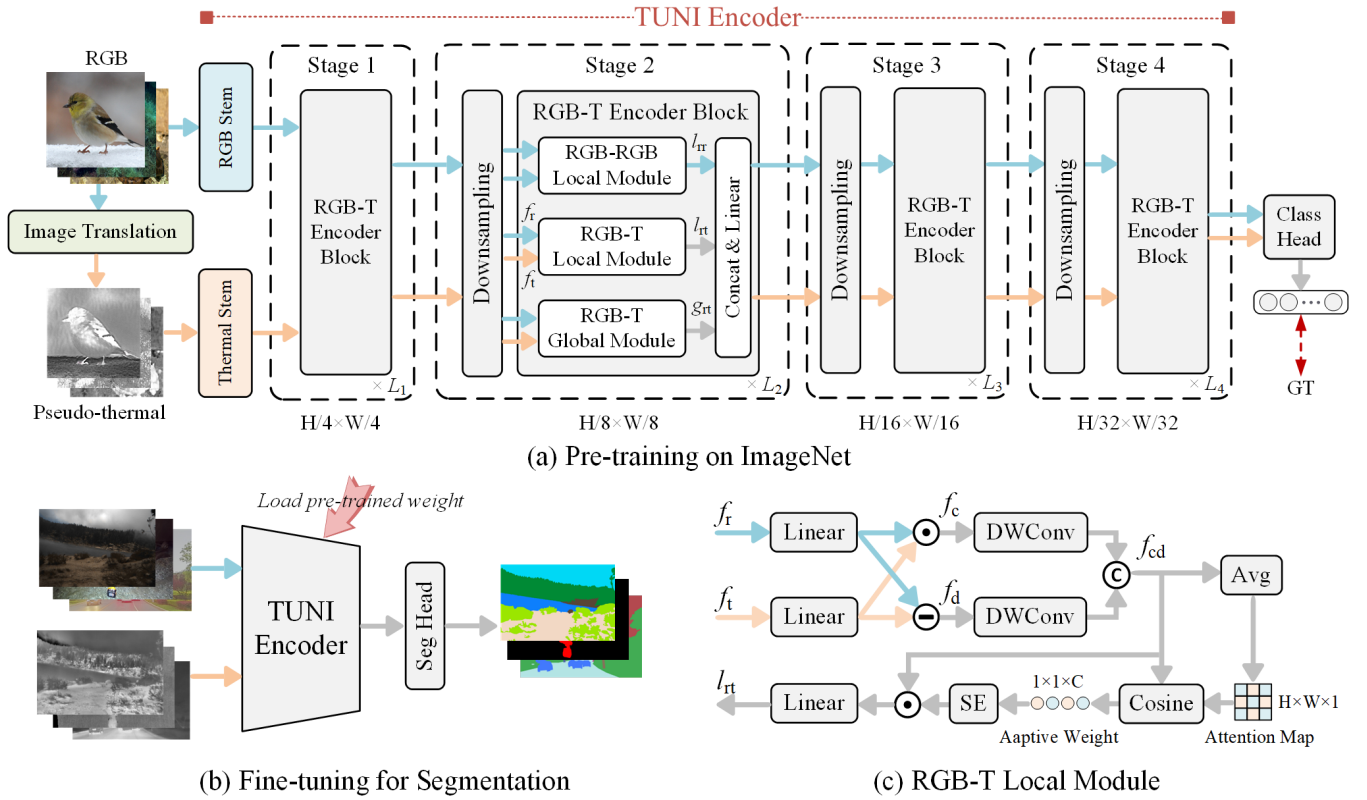


Fig. 2. Illustration of TUNI. The TUNI encoder consists of multiple stacked RGB-T encoder blocks, each of which includes an RGB-RGB local module, an RGB-T local module, and an RGB-T global module. The encoder is first pre-trained on RGB and pseudo-thermal data, and then fine-tuned with a lightweight segmentation head on downstream task datasets.

Inspired by this, we propose TUNI, which incorporates RGB-T pre-training and introduces a refined RGB-T local module, achieving efficient RGB-T semantic segmentation.

III. PROPOSED METHOD

A. Overview

The overall architecture of TUNI is illustrated in Fig. 2. It features a typical encoder–decoder architecture, with an encoder for multi-modal feature extraction and cross-modal feature fusion, and a task-specific decoder. The TUNI encoder consists of multiple blocks, each comprising three sub-modules: RGB-RGB local module, RGB-T global module and RGB-T local module. These blocks are organized into four stages $\{L_1, L_2, L_3, L_4\}$, where L_n indicates the number of blocks. Through downsampling, the feature resolutions at each stage are $H/4 \times W/4$, $H/8 \times W/8$, $H/16 \times W/16$, and $H/32 \times W/32$, where H and W denote the height and width of the original input image.

The training of TUNI consists of pre-training and fine-tuning phases. To obtain aligned thermal data, an image translation model [13] is utilized to generalize pseudo-thermal images. The TUNI encoder is pre-trained using pairs of RGB and pseudo-thermal images, along with a classification head, as shown in Fig. 2 (a). The fine-tuning is conducted on various segmentation datasets, as shown in Fig. 2 (b). Since the encoder intrinsically integrates cross-

modal feature fusion, no supplementary fusion modules are needed during fine-tuning, with only an MLP decoder [11] being employed for segmentation. Such a design enhances the model’s compactness.

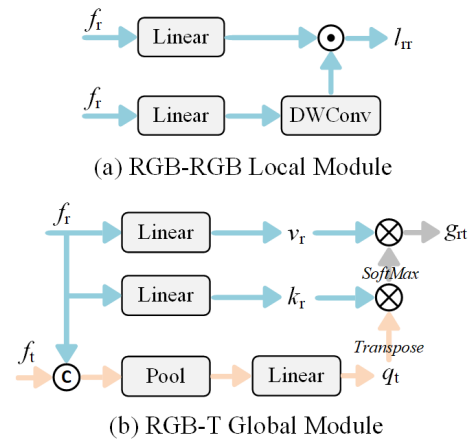


Fig. 3. Two sub-modules in the TUNI encoder: (a) RGB-RGB local module, (b) RGB-T global module.

B. RGB-RGB Local and RGB-T Global Modules

Inherited from DFormer, the RGB-RGB local and RGB-T global modules implement local RGB feature extraction

and global RGB-T feature fusion, respectively. The internal structure of the RGB-RGB local module is shown in Fig. 3 (a) (the normalization layer is omitted for convenience). In addition to basic liner layers, a depthwise separable convolution layer is employed to further extract local features. This module can be formulated as:

$$l_{rr} = \text{Linear}(f_r) \odot \text{DWConv}(\text{Linear}(f_r)) \quad (1)$$

where f_r and l_{rr} denote RGB and local-enhanced RGB features, respectively. $\text{Linear}(\cdot)$ and $\text{DWConv}(\cdot)$ denote the linear and depthwise separable convolution layers, respectively. \odot denotes the Hamilton product.

The RGB-T global module constructs cross-modal long-range dependencies via the cross-attention mechanism. As shown in Fig. 3 (b), it first obtains the value and key vectors of the RGB features, denoted as v_r and k_r , respectively:

$$\begin{cases} v_r = \text{Linear}(f_r) \\ k_r = \text{Linear}(f_r) \end{cases} \quad (2)$$

Subsequently, the query vector q_t is obtained from f_t and f_r with average pooling operation to reduce the computational burdens:

$$q_t = \text{Linear}(\text{Pool}(\text{Cat}(f_t, f_r))) \quad (3)$$

where Pool downsamples the feature maps to 7×7 . Finally, long-range dependencies between RGB-T features are captured via self-attention, yielding the global RGB-T features g_{rt} :

$$g_{rt} = \text{SoftMax}(q_t^T \otimes k_r) \otimes v_r \quad (4)$$

where SoftMax and \otimes denote the softmax operation and matrix multiplication, respectively.

C. RGB-T Local Module

In addition to capturing long-range dependencies between RGB-T features, local feature fusion plays a crucial role in leveraging cross-modal information. In DFormer, cross-modal local feature fusion is performed simply using Hamilton product, overlooking the modality-specific differences in RGB-T information. We further take into account both the consistent and distinct information in cross-modal features, and propose the RGB-T local module that employs dynamic cosine similarity to select the most informative features, as shown in Fig. 2 (c). Specifically, we extract cross-modal consistent and distinct features, f_c and f_d , using Hamilton product and absolute difference, respectively. This can be formulated as:

$$\begin{cases} f_c = \text{Linear}(f_r) \odot \text{Linear}(f_t) \\ f_d = |\text{Linear}(f_r) - \text{Linear}(f_t)| \end{cases} \quad (5)$$

Subsequently, two depthwise separable convolutions are applied to further extract the local information from the two feature components:

$$f_{cd} = \text{Cat}(\text{DWConv}(f_c), \text{DWConv}(f_d)) \quad (6)$$

With the local information extracted, the module needs to select the most informative features of f_{cd} . The Squeeze-and-Excitation block [20], a widely adopted channel attention mechanism, provides an effective strategy for this purpose. However, obtaining adaptive weights by pooling along the spatial dimension inevitably results in local information loss. In contrast, we propose an approach that derives adaptive weights by measuring the cosine similarity between the original features and the attention map produced by channel-wise average pooling. It can be formulated as:

$$W = \text{Cosine}(f_{cd}, \text{Avg}(f_{cd})) \quad (7)$$

where Avg denotes the channel-wise average pooling that obtains the attention maps $\in \mathbb{R}^{H \times W \times 1}$; Cosine represents the cosine similarity computation that derives the adaptive weight $W \in \mathbb{R}^{1 \times 1 \times C}$. Finally, the RGB-T local feature l_{rt} is obtained:

$$l_{rt} = \text{Linear}(f_{cd} \odot \text{SE}(W)) \quad (8)$$

where $\text{SE}(\cdot)$ denotes the Squeeze-and-Excitation block. With l_{rr} , g_{rt} , and l_{rt} , the RGB-T encoder block obtains updated RGB and thermal features f'_r and f'_t :

$$\begin{cases} f'_r = \text{Linear}(\text{Cat}(l_{rr}, g_{rt}, l_{rt})) \\ f'_t = \text{Linear}(\text{Cat}(l_{rr}, g_{rt}, l_{rt})) \end{cases} \quad (9)$$

D. Loss Function

Following prior research [2], [5], an MLP decoder is used to predict the semantic segmentation maps. As cross-modal feature fusion is accomplished within each block, only the RGB features from each encoder stage are fed into the MLP decoder. This can be formulated as :

$$\text{Pre} = \text{MLP}(R_1, R_2, R_3, R_4) \quad (10)$$

where $\text{MLP}(\cdot)$ refers to the MLP decoder; $\{R_1, R_2, R_3, R_4\}$ denote the RGB features from each encoder stage. Considering the varying pixel distribution of each object in the samples, we utilize the weighted cross-entropy loss. In addition, dice loss was utilized to further emphasize the segmentation of the foreground. The loss function can be formulated as follows:

$$\text{Loss} = L_{ce}(\text{Pre}, \text{GT}) + L_{dice}(\text{Pre}, \text{GT}) \quad (11)$$

where Pre and GT denote the prediction and the ground truth (GT) images, respectively.

IV. EXPERIMENT RESULTS

A. Datasets and Evaluation Metrics

In the pre-training phase, the encoder of TUNI is trained on ImageNet; in the fine-tuning phase, TUNI is trained, evaluated, and tested on FMB, PST900, and CART, which correspond to urban, underground, and wild scenes, respectively. These diverse scene settings are adopted to assess the model's generalization ability.

The FMB dataset [21] comprises 1500 pairs RGB-T images under different illumination condition in urban scenes. It labels 14 categories, including road, sidewalk, building,

TABLE I

COMPLEXITY AND PERFORMANCE COMPARISON WITH SOTA METHODS ON FMB, PST900 AND CART. ‘-’ DENOTES MISSING INFORMATION AND NON-PUBLIC MODELS, AND ‘×’ DENOTES INCOMPATIBILITY WITH THE TENSORRT ENGINE.

Model	Publication	Backbone	Params(G)↓	FLOPs(G)↓	FPS↑			mIoU↑		
					RTX 4090	Jetson Orin NX		FMB	PST900	CART
						CUDA	TensorRT			
GMNet	TIP’21	ResNet101	191.24	195.37	41	4	×	49.2	84.1	72.7
CMX	TITS’23	Mit-B2	66.57	67.20	63	3	13	61.1	84.9	74.0
CMNext	ICCV’23	Mit-B2	58.68	68.70	67	3	11	59.0	83.9	72.1
CLNet-T	KBS’24	Mit-B4	130.84	217.85	28	1	×	60.4	80.8	73.7
DFormer	ICLR’24	DFormer-B	29.5	41.9	69	5	15	61.2	85.4	73.9
MiLNet	TIP’25	Mit-B3	92.29	136.32	25	2	×	61.8	85.1	74.6
AGFNet	TITS’25	ResNet50	72.19	219.11	-	-	-	60.5	84.8	-
DFormerV2	CVPR’25	DFormerV2-S	26.70	33.90	33	2	11	61.5	83.9	72.4
CPAL	TCSVT’25	Swin-T	52.70	-	-	-	-	60.9	86.8	-
TUNI	-	TUNI	10.63	17.16	120	11	27	62.4	87.3	74.7

traffic light, traffic sign, vegetation, sky, person, car, truck, bus, motorcycle, bicycle and pole. All images have a spatial resolution of 800×600 pixels. The dataset is divided to train/test set at a ratio of 8:2.

The PST900 dataset [22] comprises 894 pairs RGB-T images captured in challenging underground environments. It provides semantic segmentation annotations for 4 classes: hand drill, backpack, fire extinguisher and survivor. Both the RGB and thermal images have a spatial resolution of 1280×720 pixels. The datasets was divided into train/test set with at a ratio of 2:1.

The CART dataset [23] contains 2282 pairs of aligned RGB-T images captured in various terrains, including rivers, lakes, coastlines, deserts and forests. It provides semantic segmentation labels for 10 classes: bare ground, rocky terrain, developed structures, road, shrubs, trees, sky, water, vehicles, and person. Both the RGB and thermal images have a spatial resolution of 960×640 pixels. The dataset was randomly partitioned into train/val/test set at a 6:1:1 ratio.

We utilize the mean intersection over union (mIoU) to assess the segmentation performance. Besides, we employ floating point operations (FLOPs), the number of parameters (Params), and FPS to assess the complexity of the model.

B. Implementation Details

The Pre-training is conducted on ImageNet with 8 RTX 4090 GPUs. An RGB-T translation model is utilized to generate aligned RGB-T image pairs and a classifier head is added to the TUNI encoder to build the classification model. The training is optimized using AdamW with the initial learning rate of $1e-3$ and weight decay $5e-2$, and the batch size is set to 1024. The resolution of RGB-T image pairs are resized to 224×224 . We train the model for 300 epochs, and the model weight with the best performance on the validation set is retained.

The Fine-tuning is conducted on FMB, PST900 and CART, with an RTX 4090 GPU. A lightweight segmentation head is added to the TUNI encoder to build the segmentation

model. Before training, the TUNI encoder is initialized with the pre-trained weights obtained from the pre-training phase, whereas the segmentation head is initialized randomly. Image resizing, random cropping and flipping were used to augment the data during training, as the configure in the original benchmark [21]–[23]. The training is optimized using Ranger with the weight decay $5e-4$ and initial learning rate $1e-4$, multiplied by $(1 - \frac{iter}{max.iter})^{power}$ during training, with a power of 0.9. For FMB, PST900, and CART, the training epochs are set to 80, 200, and 300, with corresponding batch sizes of 8, 2, and 8, respectively.

C. Comparison with SOTA models

1) Quantitative evaluation. The TUNI is compared to 9 SOTA models: GMNet [3], CMX [5], CMNeXt [17], CLNet-T [4], DFormer [14], MiLNet [18], AGFNet [19], DFormerV2 [24], CPAL [8]. As shown in Table I, high performance models are often accompanied by high complexity (e.g., MiLNet and CPAL), while lightweight models face performance bottlenecks (e.g., DFormer and DFormerV2). On the one hand, this is attributed to natural laws; on the other hand, it stems from model redundancy as well as insufficient multi-modal feature extraction and cross-modal feature fusion, as discussed in Section I. In contrast, TUNI achieves competitive performance with the best-performing SOTA models, while maintaining a lightweight design. Compared to MiLNet, TUNI achieves competitive performance while using only 1/9 of the Params, 1/8 of the FLOPs, and achieving 5× the FPS. Compared to DFormer, TUNI attains mIoU gains of 1.2%, 1.9%, and 0.8% on FMB, PST900, and CART, respectively, while further reducing model size and doubling inference speed.

To further demonstrate the model’s practical deployment capability, we evaluate its inference speed on a Jetson Orin NX. It can be observed that although SOTA models can achieve high inference speeds on an RTX 4090 GPU (e.g., DFormer, CMNeXt), they struggle to meet real-time requirements on edge devices. In contrast, TUNI achieves

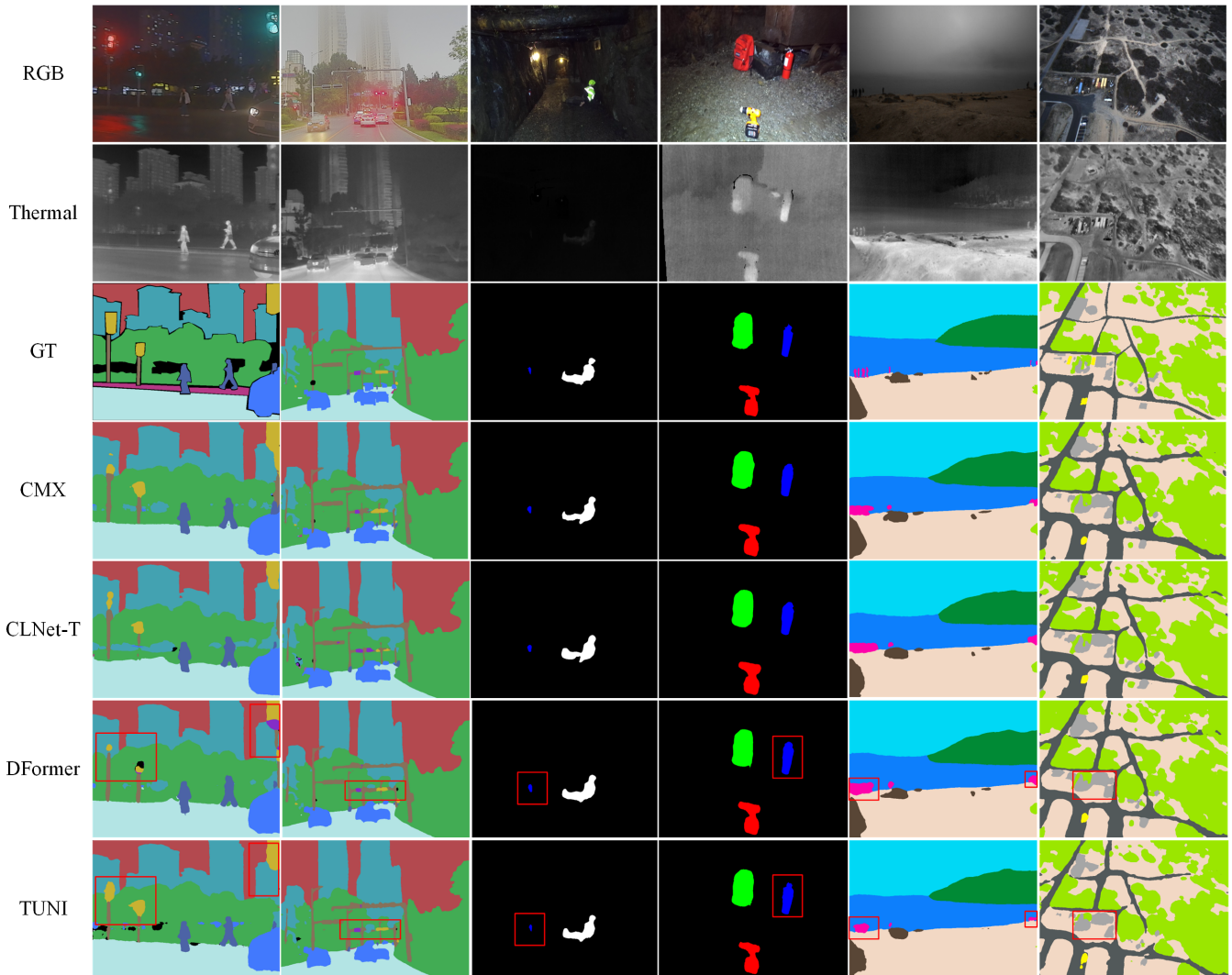


Fig. 4. Visual comparison of segmentation maps produced by TUNI and several SOTA models on FMB (left two columns), PST900 (middle two columns), and CART (right two columns).

TABLE II
COMPARISON OF DIFFERENT PRE-TRAINING STRATEGIES.

RGB-RGB	RGB-D	RGB-T	FMB	CART
✓			59.0	72.3
	✓		58.1	73.9
		✓	62.4	74.7

27 FPS with TensorRT acceleration under bf16 precision, demonstrating its real-time performance in deployment scenarios.

2) Qualitative evaluation. Fig. 4 presents the visualization results of TUNI compared with SOTA models on FMB, PST900, and CART. Although more lightweight, TUNI generates segmentation maps of a quality comparable to SOTA models, while achieving more fine-grained local results, as highlighted by the red rectangles. This may be attributed to our proposed RGB-T local module, which facilitates the

fusion of cross-modal local features.

D. Ablation Studies

To prove the effectiveness of the RGB-T pre-training and modules in TUNI encoders, we conduct a series of ablation studies.

Effectiveness of RGB-T Pre-training. Table II presents three pre-training strategies, including ‘RGB-RGB’, ‘RGB-D’, and ‘RGB-T’. ‘RGB-RGB’ denotes using identical RGB pairs as input to the TUNI encoder, ‘RGB-D’ refers to RGB-D pairs generated as in [14], and ‘RGB-T’ represents our RGB-T pre-training strategy. Compared with ‘RGB-RGB’, our approach improves mIoU by 3.4% on FMB and 2.4% on CART. This indicates that 1) the encoder pre-trained on the RGB modality is suboptimal in fine-tuning for RGB-T semantic segmentation, and 2) pre-training with RGB-T data effectively enhances the encoder’s ability to capture features from the thermal modality. Compared with ‘RGB-Depth’, our approach improves mIoU by 4.3% on FMB and 0.8% on CART. This further demonstrates that the modality

TABLE III
ABLATION OF SUB-MODULES IN TUNI ENCODER.

Module	Params	FLOPs	FMB	CART
w/o RGB-RGB Local Module	9.14	15.65	61.7	73.4
w/o RGB-T Global Module	9.52	16.43	61.3	73.3
w/o RGB-T Local Module	9.54	16.16	59.4	73.5
Modules of DFormer	10.36	16.96	59.8	73.5
TUNI	10.63	17.16	62.4	74.7

TABLE IV
ABLATION OF COMPONENTS IN THE RGB-T LOCAL MODULE.

Co&Di	SE	Cosine	Params	FLOPs	FMB	CART
			9.54	16.16	59.4	73.5
✓			10.52	17.16	61.3	74.2
✓	✓		10.63	17.16	61.6	74.2
✓	✓	✓	10.63	17.16	62.4	74.7

used during pre-training affects the performance of RGB-T semantic segmentation.

Effectiveness of Modules in the TUNI Encoder. Table III presents the ablation results of various model variants. ‘w/o RGB-RGB Local Module’ and ‘w/o RGB-T Global Module’ denote the TUNI encoder without the RGB-RGB local modules and RGB-T global modules, respectively. Compared with TUNI, their performance decreases on both datasets. ‘w/o RGB-T Local Module’ refers to the TUNI encoder with the RGB-T local modules removed, which leads to a 3.0% and 1.2% reduction in mIoU on FMB and CART, respectively. In addition, the module introduces only 1.09M parameters and 1.00G FLOPs, confirming its efficiency in fusing RGB and thermal local features. ‘Modules of DFormer’ represents the original DFormer encoder, which differs from TUNI in that it relies on a simple Hamilton product for cross-modal local feature fusion. Its performance shows little improvement compared with the variant ‘w/o RGB-T Local Module’, indicating that using a simple feature fusion strategy is ineffective for RGB-T local feature fusion. This further validates the necessity of our proposed RGB-T local module.

Effectiveness of Components in the RGB-T Local Modules. We conduct a fine-grained ablation study to assess the effectiveness of individual components, with results shown in Table IV. ‘Co&Di’ denotes the interaction of consistent and distinct cross-modal features, including linear and depthwise separable convolution layers. ‘SE’ and ‘Cosine’ denote the squeeze-and-excitation and cosine similarity components, respectively. Incorporating ‘Co&Di’ yields mIoU improvements of 1.9% and 0.7% on FMB and CART, respectively. This demonstrates the effectiveness of multi-attribute feature interaction. However, adding ‘SE’ brings little improvement, likely due to the weights obtained from global spatial averaging lacking local feature representativeness. Without introducing additional computational or parameter overhead,

incorporating cosine similarity yields mIoU gains of 0.8% and 0.5% on FMB and CART, respectively, thereby confirming the effectiveness of the proposed approach.

E. Limitation and Future Work

We adopt a simplified encoder for a lightweight, real-time model, resulting in limited performance improvement over SOTA models. Optimizing model performance via RGB-T pre-training remains an open question. In future work, we plan to progressively refine the model structure and investigate the trade-off between complexity and performance under RGB-T pre-training. While model simplification favors efficiency, knowledge distillation [25] offers a complementary approach to improve performance without significantly increasing complexity. Existing methods, however, typically apply distillation only during fine-tuning [2], [4], limiting performance due to scarce data and pre-trained backbones. We will explore distillation during pre-training and its impact on downstream tasks.

V. CONCLUSION

We propose a real-time RGB-T semantic segmentation model, TUNI, for robotic application. The TUNI encoder integrates multi-modal feature extraction and cross-modal feature fusion, with RGB-T pre-training effectively enhancing both components. In addition, an RGB-T local module is introduced to emphasize salient consistent and distinct local features across RGB-T modalities, thereby improving cross-modal local fusion. As a result, we achieve the following outcomes: 1) TUNI achieves competitive performance with SOTA models on FMB, PST900 and CART, with fewer parameters and lower computational cost, 2) TUNI demonstrates finer segmentation details in local regions, 3) TUNI achieves 120 FPS on an RTX 4090 and 27 FPS on a Jetson Orin NX, demonstrating its real-time capability in practical applications, and 4) ablation results further confirm the contributions of the RGB-T pre-training and the RGB-T local module to performance improvement. In future research, we aim to investigate the trade-off between model accuracy and efficiency, and further explore how knowledge distillation in the pre-training stage influences fine-tuning performance.

REFERENCES

- [1] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. 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)*, 2017, pp. 5108–5115.
- [2] X. Guo, T. Liu, Y. Mou, S. Chai, B. Ren, Y. Wang, W. Shi, S. Liu, and W. Zhou, “Transferring prior thermal knowledge for snowy urban scene semantic segmentation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 8, pp. 12474–12487, 2025.
- [3] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, “Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation,” *IEEE Transactions on Image Processing*, vol. 30, pp. 7790–7802, 2021.
- [4] X. Guo, W. Zhou, and T. Liu, “Contrastive learning-based knowledge distillation for rgb-thermal urban scene semantic segmentation,” *Knowledge-Based Systems*, vol. 292, p. 111588, 2024.
- [5] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelwagen, “CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers,” *IEEE Transactions on intelligent transportation systems*, vol. 24, no. 12, pp. 14679–14694, 2023.

- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.
- [7] S. Dong, Y. Feng, Q. Yang, Y. Huang, D. Liu, and H. Fan, "Efficient multimodal semantic segmentation via dual-prompt learning," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 14 196–14 203.
- [8] Y. Liu, P. Wu, M. Wang, and J. Liu, "Cpal: Cross-prompting adapter with loras for rgb+x semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 6, pp. 5858–5871, 2025.
- [9] W. Zhou, J. Hong, W. Yan, and Q. Jiang, "Modal evaluation network via knowledge distillation for no-service rail surface defect detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3930–3942, 2024.
- [10] W. Zhou, X. Sun, X. Qian, and M. Fang, "Asymmetrical contrastive learning network via knowledge distillation for no-service rail surface defect detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 7, pp. 12 469–12 482, 2025.
- [11] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.
- [12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.
- [13] D. Lee, M. Jeon, Y. Cho, and A. Kim, "Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8291–8298.
- [14] B. Yin, X. Zhang, Z.-Y. Li, L. Liu, M.-M. Cheng, and Q. Hou, "DFormer: Rethinking RGBD representation learning for semantic segmentation," in *The Twelfth International Conference on Learning Representations*, 2024.
- [15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10 002.
- [17] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen, "Delivering arbitrary-modal semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1136–1147.
- [18] J. Liu, H. Liu, X. Li, J. Ren, and X. Xu, "Milnet: Multiplex interactive learning network for rgb-t semantic segmentation," *IEEE Transactions on Image Processing*, vol. 34, pp. 1686–1699, 2025.
- [19] X. Zhou, X. Wu, L. Bao, H. Yin, Q. Jiang, and J. Zhang, "Agfnet: Adaptive gated fusion network for rgb-t semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 5, pp. 6477–6492, 2025.
- [20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [21] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 8081–8090.
- [22] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "Pst900: Rgb-thermal calibration, dataset and segmentation network," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 9441–9447.
- [23] C. Lee, M. Anderson, N. Ranganathan, X. Zuo, K. Do, G. Gkioxari, and S.-J. Chung, "Caltech aerial rgb-thermal dataset in the wild," in *Computer Vision – ECCV 2024: 18th European Conference*, 2024, p. 236–256.
- [24] B.-W. Yin, J.-L. Cao, M.-M. Cheng, and Q. Hou, "Dformerv2: Geometry self-attention for rgbd semantic segmentation," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 19 345–19 355.
- [25] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.