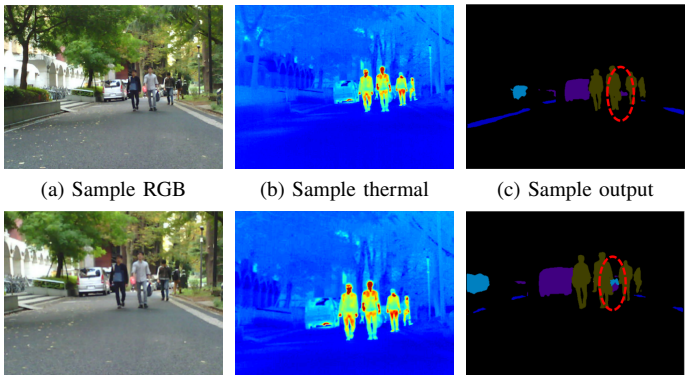
Temporal Consistency for RGB-Thermal Data-based Semantic Scene Understanding

Haotian Li, Henry K. Chu, and Yuxiang Sun

***Abstract*—Semantic scene understanding is a fundamental capability for autonomous vehicles. Under challenging lighting conditions, such as nighttime and on-coming headlights, the semantic scene understanding performance using only RGB images are usually degraded. Thermal images can provide complementary information to RGB images, so many recent semantic segmentation networks have been proposed using RGB-Thermal (RGB-T) images. However, most existing networks focus only on improving segmentation accuracy for single image frames, omitting the information consistency between consecutive frames. To provide a solution to this issue, we propose a temporal-consistent framework for RGB-T semantic segmentation, which introduces a virtual view image generation module to synthesize a virtual image for the next moment, and a consistency loss function to ensure the segmentation consistency. We also propose an evaluation metric to measure both the accuracy and consistency for semantic segmentation. Experimental results show that our framework outperforms state-of-the-art methods.**

***Index Terms*—Temporal Consistency, Multi-modal Fusion, RGB-Thermal, Semantic Segmentation, Autonomous Vehicles.**



(a) Sample RGB (b) Sample thermal (c) Sample output

(d) Synthesized RGB (e) Synthesized thermal (f) Synthesized output

Fig. 1. Inconsistent segmentation results across consecutive frames. The RGB-T images in the first row are sampled from the MFNet [8] dataset. The RGB-T images in the second row are the synthesized images at the next moment. The thermal images are visualized with the *jet* color map. We use the recent network CMX [13] to obtain the segmentation results. The blue and purple colors in the segmentation maps represent bike and car, respectively.

I. Introduction

SEMANTIC scene understanding based on semantic image segmentation is an essential capability for autonomous vehicles. It provides fundamental perceptual information for downstream tasks, such as localization [1–3] and autonomous navigation [4–6]. Most existing semantic segmentation networks are designed with RGB images from visible cameras. Due to the intrinsic limitations of visible cameras, the performance of these networks may be degraded under challenging lighting conditions, such as nighttime, glares, and on-coming headlights. Recently, semantic segmentation based on RGB-Thermal (RGB-T) images has been proposed to address this issue [7], since thermal imaging cameras do not use visible lights for imaging and thermal images can provide complementary information to RGB images. Research progress has been made using convolutional neural network (CNN) [8–11] and Transformer [12–14].

However, most RGB-T semantic segmentation networks primarily focus on enhancing segmentation accuracy solely for single image frames, overlooking the segmentation consistency between consecutive frames [8, 9, 13, 14]. Fig. 1 shows the segmentation results by a recent network CMX [13] for a sample RGB-T image from the MFNet dataset [8], and a synthesized RGB-T image at the next moment generated by our virtual view image generation (VVIG) module. We can observe the degraded segmentation performance across consecutive frames. For example, the segmentation result for the synthesized RGB-T image wrongly classifies the car as a bike, resulting in inconsistent segmentation across consecutive frames (highlighted by the red ellipses). The inconsistent segmentation results are not expected by most downstream tasks. To improve the segmentation consistency across consecutive frames, some works on semantic segmentation have tried to use optical flow [15, 16]. But in real-world applications, especially under challenging lighting conditions, it is difficult to compute accurate optical flow.

To provide a solution to this issue, we propose a temporal-consistent framework to improve the segmentation consistency for RGB-T semantic segmentation. We design a loss function in this framework to ensure consistency for the segmentation results across different frames. We also introduce a new evaluation metric to measure both consistency and accuracy for semantic segmentation. The main contributions of this letter are summarized as follows:

- We design a novel temporal-consistent framework for RGB-T semantic segmentation, including a new method to synthesize images at the next moment. Our code is open-

Manuscript received February 23, 2024; Revised June 3, 2024; Accepted August 21, 2024. This paper was recommended for publication by Editor Hyungpil Moon upon evaluation of the Associate Editor and Reviewers' comments. This work was supported in part by Hong Kong Innovation and Technology Fund under Grant ITS/145/21, and in part by City University of Hong Kong under Grant 9610675. *(Corresponding author: Yuxiang Sun.)*

Haotian Li and Henry K. Chu are with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (e-mail: haotian.li@connect.polyu.hk; henry.chu@polyu.edu.hk).

Yuxiang Sun is with the Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong (email: yx.sun@cityu.edu.hk, sun.yuxiang@outlook.com).

Digital Object Identifier (DOI): see top of this page.

sourced[1].
- We design a novel loss function to ensure segmentation consistency across different frames.
- We design a new evaluation metric to measure both accuracy and consistency for semantic segmentation.

II. Related Work

A. RGB-T Semantic Segmentation

RTFNet [9] uses the two-encoders-one-decoder fusion structure to fuse RGB and thermal images. CACFNet [17] utilizes cross-modal attention and cascaded fusion to enhance RGB-T feature complementarity. MMSMCNet [18] uses modal memory fusion and morphological multi-scale assistance to enhance cross-modal features. Liang et al. [19] proposed the Explicit Attention-Enhanced Fusion (EAEF), which adapts to different cases of RGB-T data availability. Lv et al. [20] introduced CAINet, which leverages auxiliary tasks and global context to enhance the complementary reasoning and detailed aggregation of multi-modal features. Dong et al. [21] proposed EGFNet that uses prior edge maps and multi-modal fusion modules to enhance the feature maps. Inspired by Vision Transformer (ViT) [22] and Segmentation Transformer (SETR) [23], several transformer-based methods [13, 14, 24] have been applied to RGB-T semantic segmentation. Zhang et al. [13] introduced CMX, an extension of Segformer [12] to multi-modal tasks.

B. Segmentation Consistency

To ensure temporal consistency for semantic segmentation, Cheng et al. [15] proposed a bi-directional framework to obtain the foreground segmentation and optical flow at the same time. The optical flow is used as the complementary information for the segmentation task. Nilsson et al. [16] applied optical flow for video segmentation. They enhanced segmentation accuracy and consistency by utilizing the unlabeled RGB images in the dataset and propagating labels through optical flow. Zhang et al. [25] proposed AuxAdapt to improve the temporal consistency of RGB segmentation networks without using optical flow, by learning from the own decisions of the network and a small auxiliary network.

C. Evaluation Metric for Consistency

The widely used intersection-over-union (IoU) is the standard evaluation metric for measuring segmentation accuracy. To evaluate segmentation consistency, Liu et al. [26] used optical flow to warp the segmentation map of the current frame to align with the previous frame, and computed IoU between the warped segmentation map and the original segmentation map of the previous frame. They define this as the temporal consistency (TC) score. Zhang et al. [25] extended the TC metric to evaluate the segmentation consistency. Park et al. [27] proposed a perceptual-consistency-based metric, which calculates the temporal consistency by comparing the average cosine similarity of the feature maps of consecutive frames.

[1]https://github.com/lab-sun/Temporal-Consistent-RGBT-Segmentation

D. Difference from Existing Work

The existing works on RGB-T semantic segmentation focus on improving segmentation accuracy for single image frames. We use the information of consecutive frames to enhance the segmentation consistency and further improve the segmentation accuracy.

The existing consistency evaluation metrics only measure the consistency of segmentation results, regardless of the segmentation correctness. So, these metrics may fail to evaluate the performance when the segmentation results of consecutive frames produce the same errors for the same object. We introduce a novel metric that takes into account the ground truth of the segmentation result. The metric evaluates both the accuracy and the consistency of a segmentation network.

III. The Proposed Method

A. The Framework Overview

To study segmentation consistency, we need datasets with temporal-sequential images. But the existing RGB-T dataset, such as the MFNet dataset [8], includes only discrete image frames. So, we propose the virtual view image generation (VVIG) module to synthesize the sequential frames based on the existing dataset. The overview of our proposed temporal-consistent framework is shown in Fig. 2. Img_1, which represents the current frame, is the real image sampled from the MFNet [8]. We use the VVIG module to synthesize Img_2, which is the frame at the next moment. We choose CMX [13] as the segmentation network. It utilizes a unified fusion approach with a Cross-Modal Feature Rectification Module (CM-FRM) to calibrate bi-modal features. Additionally, a Feature Fusion Module (FFM) for long-range context exchange, achieving state-of-the-art performance across various RGB-X modalities [13]. In this work, the encoders and decoders for the two frames are the same. They share the same weights. The CMX network can be replaced with other RGB-T semantic segmentation networks.

B. Virtual View Image Generation Module

As aforementioned, we propose the VVIG module to synthesize the image at the next moment. Fig. 3 shows the pipeline. To mimic the camera's viewpoint at the next moment, we need to simulate both its rotation and translation. For rotation, we randomly generate the Euler angles α, γ, and β about the $\hat{Z}$, $\hat{Y}$, and $\hat{X}$ axes in the image coordinate system. The ranges of the angles are: $\alpha \in [-5°, 5°]$, $\gamma \in [-10°, 10°]$, $\beta \in [-10°, 10°]$. We use the Euler angles and the camera intrinsic matrix to calculate the virtual view transformation (VVT) matrix, which is then used to transform the original image. M_{VVT} is found by:

$$M_{\mathrm{VVT}} = K R_z(\alpha) R_y(\beta) R_x(\gamma) K^{-1}, \tag{1}$$

where K represents the camera intrinsic matrix. The VVT matrix becomes equivalent to a homography matrix when the camera motion involves only rotations, without any translations. Detailed derivation and descriptions are presented in the Appendix.

After simulating rotation with the VVT matrix and obtaining the distorted image, we discard the parts that appear beyond

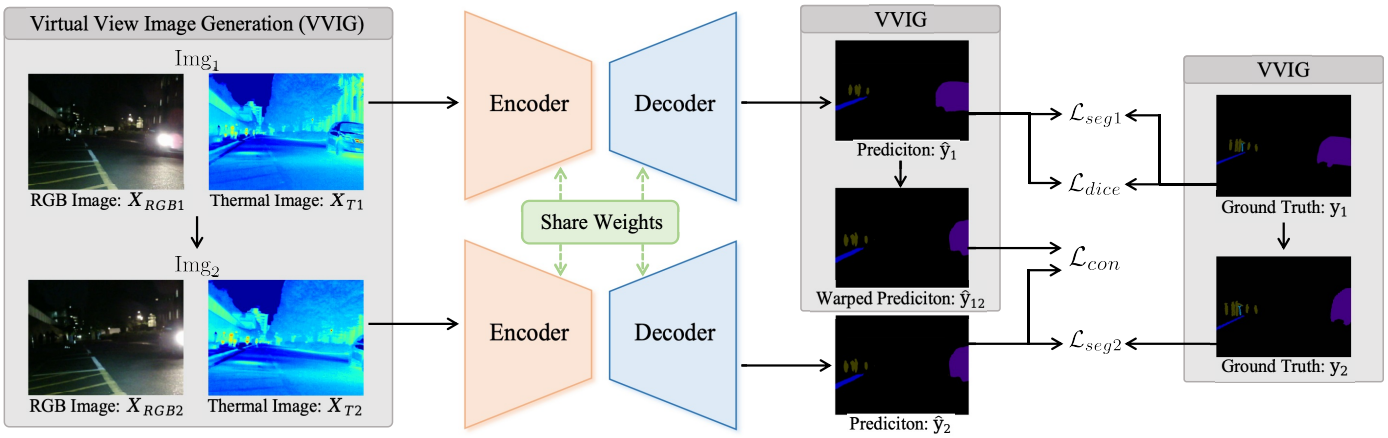


Fig. 2. The overview of our temporal-consistent framework for RGB-T semantic segmentation. Img_1 is sampled from MFNet. Img_2 is generated using our virtual view image generation (VVIG) module. The encoders and decoders are borrowed from CMX [13]. The warped prediction $\hat{y}_{12}$ and ground truth y_2 are also generated by the VVIG module. The *jet* color map is used here to visualize the thermal images.

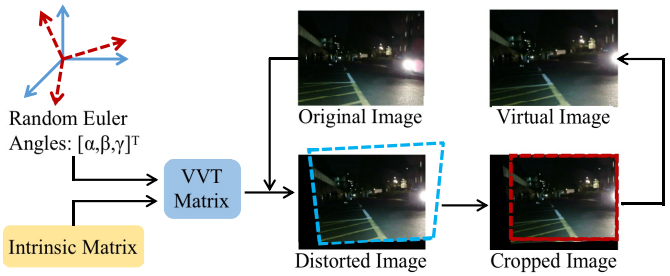


Fig. 3. The pipeline of our VVIG module. α, γ, β are respectively the randomly generated Euler angles about the $\hat{Z}$, $\hat{Y}$, $\hat{X}$ axes. The VVT matrix (i.e., M_{VVT}) is generated from the Euler angles and the intrinsic matrix. The blue dashed box on the transformed image represents the shape of the original image after the M_{VVT} transformation. The red dashed box on the cropped image shows the largest inner rectangle within the valid range for cropping. The virtual image is obtained from the cropped image by interpolation.

the field-of-view of the image, indicated by the white areas in the blue dashed box. The remaining area within the blue dashed box is then cropped using the red dashed box to its largest inner rectangle, defining the effective pixel area of the virtual image. Finally, the cropped image is resized to match the original image's resolution by using interpolation to produce the virtual image. These cropping and resizing operations effectively simulate the forward movement of the camera. To maximize the protection of edge information in objects, we utilize bilinear interpolation for RGB and thermal images, and nearest-neighbor difference for ground truth during warping and resizing operations. As shown in Fig. 2, we feed an RGB image, a thermal image, a ground-truth image, and the prediction of Img_1 into the VVIG module. The module then generates the corresponding Img_2.

C. The Loss Functions

As shown in Fig. 2, the framework conducts semantic segmentation on Img_1 and Img_2 separately. We adopt the cross-entropy loss (i.e., $\mathcal{L}_{seg1}$ and $\mathcal{L}_{seg2}$) for the semantic segmentation task. In addition, we apply the Dice loss [28] (i.e., $\mathcal{L}_{dice}$) to further improve the segmentation accuracy.

We propose two consistency loss functions (i.e., $\mathcal{L}_{con}$ and $\mathcal{L}_{con-acc}$) to improve the segmentation consistency across consecutive frames. Specifically, $\mathcal{L}_{con}$ transforms the segmentation map of Img_1 to Img_2 through M_{VVT}, leading to $\hat{y}_{12}$, which is then compared with the segmentation map of Img_2, $\hat{y}_2$. The greater the similarity between $\hat{y}_{12}$ and $\hat{y}_2$, the better the segmentation consistency.

First, we have to transform the segmentation map of the current frame to the position of next frame by $\hat{y}_{12} = M_{\text{VVT}}\hat{y}_1$, where M_{VVT} is used as the true value. Then, $\mathcal{L}_{con}$ is calculated to measure the inconsistency between the consecutive frames:

$$\mathcal{L}_{con} = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{h=1,w=1}^{H\times W}\hat{y}_{12}^{(c,h,w)}\cdot\hat{y}_{2}^{(c,h,w)}}{\sum_{h=1,w=1}^{H\times W}(\hat{y}_{12}^{(c,h,w)}+\hat{y}_{2}^{(c,h,w)}+\sigma)}, \quad (2)$$

where h, w and c denote the row, column and class indices of predictions, respectively. $H \times W$ denotes the number of pixels and C denotes the number of classes. σ is a very small positive number that prevents the denominator from being zero. In this paper, we set $\sigma = 1 \times 10^{-7}$.

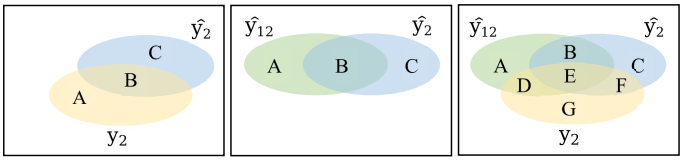
Considering that $\mathcal{L}_{con}$ only constrains the consistency between the segmentation maps of consecutive frames, if the segmentation maps make the same incorrect prediction for the same pixel, the loss function cannot constrain them effectively. So, we propose $\mathcal{L}_{con-acc}$ based on $\mathcal{L}_{con}$. This loss function incorporates the ground truth of the segmentation map of Img_2, y_2, and measures the consistency among $\hat{y}_{12}$, $\hat{y}_2$, and y_2:

$$\mathcal{L}_{con-acc} = 1 - \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{h=1,w=1}^{H\times W}\hat{y}_{12}^{(c,h,w)}\cdot\hat{y}_{2}^{(c,h,w)}\cdot y_{2}^{(c,h,w)}}{\sum_{h=1,w=1}^{H\times W}(\hat{y}_{12}^{(c,h,w)}+\hat{y}_{2}^{(c,h,w)}+y_{2}^{(c,h,w)}+\sigma)}, \quad (3)$$

where y_2 is used as a mask to improve the segmentation accuracy.

D. The Evaluation Metric

Intersection-over-Union (IoU) is a widely-used metric that evaluates segmentation accuracy for a single frame. Temporal consistency (TC) [26] is proposed to evaluate the consistency of RGB segmentation across consecutive frames. While TC can capture the segmentation consistency for consecutive frames, it fails to account for scenarios where consecutive frames make the same incorrect prediction for the same target.



(a) IoU (b) TC (c) CA

Fig. 4. (a) Intersection over union (IoU): $y_2 = A + B$, represents the ground truth of Img_2; $\hat{y}_2 = B + C$, represents the prediction of Img_2. (b) Temporal consistency (TC): $\hat{y}_{12} = A + B$, represents the prediction of Img_1 mapped to Img_2 by M_{VVT}; $\hat{y}_2 = B + C$, represents the prediction of Img_2. (c) Consistent accuracy (CA): $\hat{y}_{12} = A+B+D+E$, represents the prediction of Img_1 mapped to Img_2 by M_{VVT}; $\hat{y}_2 = B + C + E + F$, represents the prediction of Img_2; $y_2 = D + E + F + G$, represents the ground truth of Img_2.

To address this issue, we propose consistent accuracy (CA), which evaluates segmentation results in terms of both consistency and accuracy. Fig. 4 shows the schematic diagrams of IoU, TC and CA. According to the definition of IoU, it is calculated as $IoU = \frac{B}{A+B+C} \times 100\%$, indicating the similarity between the prediction and the ground truth. Liu et al. [26] proposed TC to measure the segmentation consistency of RGB images. As shown in Fig. 4(b), TC is defined as $TC = \frac{B}{A+B+C} \times 100\%$, which measures the similarity between the current frame, $\hat{y}_2$, and the warped previous frame, $\hat{y}_{12}$, by optical flow. As aforementioned, it is challenging to calculate optical flow under unsatisfactory lighting conditions. So, we warp the current frame using M_{VVT} and utilize the ground truth y_2 as a supervision signal to enhance the segmentation accuracy. As shown in Fig. 4(c), CA is defined as $CA = \frac{E}{A+B+C+D+E+F+G} \times 100\%$. $CA = 100\%$ only if the predictions of consecutive frames are consistent and match the ground-truth labels. Otherwise, CA is decreased by both inconsistency and inaccuracy.

IV. Experimental Results and Discussions

A. Datasets

We used the MFNet dataset [8] for our experiments. The dataset consists of 2,390 pairs of RGB-T images. It has 9 classes (i.e., unlabelled background, car, person, bike, curve, car stop, guardrail, color cone, and bump). We used the same split scheme as [8] to train our network, that is, 1568 pairs for training, 392 pairs for validation, and 393 pairs for testing.

B. Training Details

We implement our proposed method in PyTorch and train the networks with an NVIDIA RTX 3090 (24GB RAM) graphics card. We employ the CMX [13] network, utilizing the same dual-stream encoder to extract features from RGB and thermal modalities, and a decoder that integrates these features. Specifically, we adopt the four-stage Mix Transformer (MiT) encoder, pre-trained on ImageNet [29]. This hierarchically structured Transformer encoder can avoid interpolation of positional codes and hence obtain multi-scale features. We choose MiT-B2 as the backbone to trade-off the performance and computational expenses. The decoder is a multi-layer perceptron (MLP) with an embedding dimension of 512, as proposed in SegFormer [12]. We train the network with the AdamW optimizer, using a weight decay rate of 1×10^{-3}. The initial learning rate is 6×10^{-5}, and we use the poly learning rate schedule [30]. We use mean precision (mPre), mean accuracy (mAcc), mean F1 (mF1), mean intersection over union (mIoU), TC and CA for the quantitative evaluations.

TABLE I

Results (%) of the ablation study on the VVIG module with different losses. Img_1 and Img_2 indicate the losses applied to the current frame and next frame, respectively. The best results are highlighted in bold font. We use the average results generated from the same 3 sets of random Euler angles as the final results of TC (%) and CA (%).

No.	Losses		mPre	mAcc	mF1	mIoU	TC	CA
	Img_1	Img_2						
(A)	$\mathcal{L}_{seg1}$	—	75.93	67.32	69.85	58.19	34.58	22.29
(B)	$\mathcal{L}_{seg1}$	$\mathcal{L}_{seg2}$	75.41	70.37	70.06	58.76	34.52	22.34
(C)	$\mathcal{L}_{seg1}$	$\mathcal{L}_{seg2}$+$\mathcal{L}_{dice2}$	74.91	73.86	71.21	59.47	34.69	22.74
(D)	$\mathcal{L}_{seg1}$	$\mathcal{L}_{seg2}$+$\mathcal{L}_{con}$	74.66	70.07	70.57	58.84	35.86	22.89
(E)	$\mathcal{L}_{seg1}$	$\mathcal{L}_{seg2}$+$\mathcal{L}_{con-acc}$	74.03	69.74	70.15	59.02	34.58	22.61
(F)	$\mathcal{L}_{seg1}$+$\mathcal{L}_{dice1}$	$\mathcal{L}_{seg2}$	72.90	**74.53**	72.47	60.33	34.30	22.70
(G)	$\mathcal{L}_{seg1}$+$\mathcal{L}_{dice1}$	$\mathcal{L}_{seg2}$+$\mathcal{L}_{dice2}$	**76.93**	71.04	**72.99**	**60.78**	32.35	22.65
(H)	$\mathcal{L}_{seg1}$+$\mathcal{L}_{dice1}$	$\mathcal{L}_{seg2}$+$\mathcal{L}_{con}$	74.01	74.17	71.46	59.90	**36.40**	**23.29**
(I)	$\mathcal{L}_{seg1}$+$\mathcal{L}_{dice1}$	$\mathcal{L}_{seg2}$+$\mathcal{L}_{con-acc}$	74.49	72.26	71.75	60.01	34.41	22.77

TABLE II

Results (%) of the ablation study on the components of the VVIG module. Rot. and Tra. indicate the rotation and translation of the VVIG module, respectively. The best results are highlighted in bold font. We use the average results generated from the same 3 sets of random Euler angles as the final results of TC (%) and CA (%).

No.	VVIG		mPre	mAcc	mF1	mIoU	TC	CA
	Rot.	Tra.						
(A)			**75.93**	67.32	69.85	58.19	34.58	22.29
(B)	✓		73.47	**74.87**	**71.55**	59.75	34.58	22.61
(C)		✓	70.93	74.50	70.74	58.89	35.47	22.71
(D)	✓	✓	74.01	74.17	71.46	**59.90**	**36.40**	**23.29**

C. Ablation Study

1) Ablation on the VVIG Module: To demonstrate the effectiveness of the VVIG module, we use CMX [13] based on MiT-B2 as the baseline, denoted as variant (A) in Tab. I. Variant (A) only uses the cross-entropy loss, which is the same as [13]. As shown in Tab. I, variants (B) to (I) all employ the VVIG module to generate the virtual view images to improve the segmentation consistency. To improve the segmentation accuracy, we use the Dice loss $\mathcal{L}_{dice}$ and two cross-entropy losses, $\mathcal{L}_{seg1}$ and $\mathcal{L}_{seg2}$ for the segmentation of Img_1 and Img_2, respectively. According to the definitions given in Eq. (2) and Eq. (3), we use $\mathcal{L}_{con}$ and $\mathcal{L}_{con-acc}$ to represent our proposed consistency losses for the segmentation of Img_2.

From Tab. I, despite achieving around 60% mIoU, all of the variants have a low TC consistency less than 40%. Our proposed consistency evaluation metric CA is even lower than 25%. This shows that existing methods, despite having good accuracy for single frames, suffer from poor consistency. From the results of variants (B) to (I), we find that the VVIG module can enhance the segmentation accuracy and consistency, resulting in higher mF1, mIoU, TC and CA than those of variant (A). The results of variants (C) and (F) show that the Dice loss can greatly enhance the segmentation accuracy, achieving much higher mF1 and mIoU. However, despite the significant improvement in mIoU for variant (F)

over variant (A), the TC decreases. The reason may be that TC only measures segmentation consistency between consecutive frames without considering segmentation accuracy. On the other hand, our proposed CA follows the same trend as mIoU, since it incorporates the ground truth of the segmentation map as a supervision signal, and the CA reaches 100% only when the segmentation results of consecutive frames are perfectly accurate and consistent.

By comparing the results of variants (C), (D) and (E), we find that $\mathcal{L}_{dice}$ greatly enhances the segmentation accuracy, $\mathcal{L}_{con}$ greatly enhances the segmentation consistency, and $\mathcal{L}_{con-acc}$ achieves a balance between segmentation accuracy and consistency. From variant (F), we use $\mathcal{L}_{dice}$, $\mathcal{L}_{con}$ and $\mathcal{L}_{con-acc}$ on Img_2, resulting in variants (G), (H) and (I), respectively. Although variant (G) has the highest mIoU, its TC and CA are the lowest. Notably, TC of variant (G) is even lower than that of variant (A), indicating that $\mathcal{L}_{dice}$ reduces segmentation consistency despite its improvement in segmentation accuracy. While the mIoU of variant (H) is slightly lower than that of variant (G), it shows a significant improvement over that of the baseline variant (A). Importantly, variant (H) achieves the highest TC and CA, demonstrating that $\mathcal{L}_{con}$ can substantially enhance segmentation consistency while improving segmentation accuracy. Compared to variant (H), variant (I) achieves higher segmentation accuracy but lower segmentation consistency. Variant (I) represents a compromise between variants (G) and (H), as it incorporates the ground truth of the segmentation map as a constraint in $\mathcal{L}_{con-acc}$. Our experimental results and analyses indicate that, in semantic segmentation task, there is an inherent trade-off between segmentation accuracy and consistency. The above experiments demonstrate, emphasizing more on the segmentation accuracy would hinder consistency to some degree. Segmentation consistency is essential in practical applications. So, in this work, we adopt segmentation consistency as an overall evaluation metric, by which variant (H) is the best method.

To better analyze the impact of simulating rotation and translation in the VVIG module, an ablation study is conducted, the results of which are displayed in Tab. II. Variant (A) represents the results without the VVIG module, variant (B) represents the results with the VVIG module simulating only rotation, variant (C) represents the results with the VVIG module simulating only translation, and variant (D) represents the results using the VVIG module that simulates both rotation and translation. The results indicate that simulating rotation significantly enhances mIoU, thereby improving segmentation accuracy. Meanwhile, simulating translation notably improves TC and CA, thereby enhancing segmentation consistency. Simultaneously simulating both rotation and translation leads to great improvements in both segmentation accuracy and consistency, thereby demonstrating the importance of simulating both rotation and translation in the VVIG module.

2) Ablation on Evaluation Metric: To further analyze the segmentation consistency performance of our proposed variants (G), (H), and (I), we conduct a comparative analysis in terms of the TC and CA across various classes within the MFNet dataset [8], as depicted in Fig. 5. TC measures the temporal consistency of the semantic segmentation results between consecutive frames. As shown in Fig. 5, variant (H) achieves the highest TC for most classes, except for person, where variant (I) performs better. It is noteworthy that variant (H) shows a significant improvement in TC for the car stop and guardrail compared to other variants, indicating that it has a greater effect on the segmentation consistency of challenging objects.

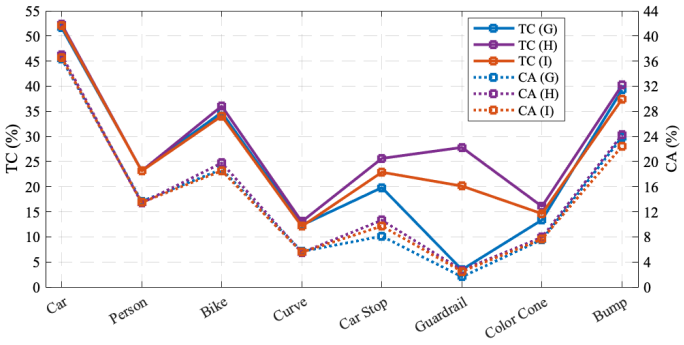


Fig. 5. The ablation study results (%) of temporal consistency (TC) and consistent accuracy (CA). The horizontal axis shows each class in the MFNet dataset [8]. Variants (G), (H) and (I) are the three variants of our method. The figure illustrates the superiority of variant (H) in terms of segmentation consistency. The figure is best viewed in color.

TABLE III
RESULTS (%) OF THE ABLATION STUDY ON THE COEFFICIENTS OF THE LOSS FUNCTIONS. α AND β INDICATE THE IMPORTANCE OF DICE LOSS AND CONSISTENCY LOSS, RESPECTIVELY. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT. WE USE THE AVERAGE RESULTS GENERATED FROM THE SAME 3 SETS OF RANDOM EULER ANGLES AS THE FINAL RESULTS OF TC (%) AND CA (%).

No.	Coefficients		mPre	mAcc	mF1	mIoU	TC	CA
	α	β						
(A)	0.5	0.5	73.70	73.07	71.27	59.56	35.02	22.87
(B)	0.5	1.0	73.20	71.26	70.09	58.63	34.94	22.53
(C)	0.5	2.0	71.26	74.81	71.47	59.12	36.36	23.06
(D)	1.0	0.5	71.37	**77.53**	**71.78**	59.89	35.22	22.76
(E)	1.0	1.0	**74.01**	74.17	71.46	**59.90**	**36.40**	**23.29**
(F)	1.0	2.0	71.44	74.51	70.76	58.74	36.31	23.10
(G)	2.0	0.5	72.99	72.54	70.47	58.85	34.75	22.49
(H)	2.0	1.0	73.58	73.64	71.48	59.85	35.70	23.01
(I)	2.0	2.0	70.23	73.00	69.90	58.20	36.16	22.84

Contrary to the TC metric, the CA metric employs the ground truth of segmentation map as the supervision signal. This approach ensures that, in scenarios where segmentation results across consecutive frames are consistent but inaccurate, CA does not yield relatively high values as observed with TC. This distinction is exemplified in the results for the car stop and guardrail depicted in Fig. 5. Despite the equal ranking of the three variants in TC and CA for both car stop and guardrail, the gap between the three in TC is significantly larger than that in CA. This implies that evaluating solely the consistency of segmentation results introduces considerable uncertainty, particularly when the segmentation results of consecutive frames exhibit the same error for a specific object. So, by leveraging the truth value of segmentation map, CA can more accurately reflect the segmentation performance. Variant (H) achieves the highest CA in almost all classes, indicating that its segmentation results are both accurate and consistent.

3) Ablation on Coefficients of Loss Functions: From the results in Tab. I, we select variant (H) as the best method. It uses the cross-entropy loss and Dice loss on Img_1, and cross-entropy

TABLE IV
THE COMPARATIVE PER-CLASS RESULTS ON THE MFNET DATASET. WE USE ACC (%) AND IOU (%) FOR EACH CLASS AND THE MACC (%) AND MIOU (%) FOR ALL THE CLASSES. THE RESULTS DEMONSTRATE THE SUPERIORITY OF OUR METHOD, WITH THE TOP TWO RESULTS IN EACH COLUMN HIGHLIGHTED IN **BOLD** AND UNDERLINE. THE PUBLICATION VENUE IS FOLLOWED BY THE PUBLICATION YEAR. SINCE CMX WITH MIT-B4 BACKBONE HAS NOT RELEASED THE ACC RESULTS BASED ON ITS PRE-TRAINED WEIGHTS, WE USE "-" TO INDICATE THE DATA ABSENCE.

Method	Backbone	Venue	Car		Person		Bike		Curve		Car Stop		Guardrail		Color Cone		Bump		mAcc	mIoU
			Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
RTFNet [9]	ResNet-152	RAL'19	93.0	87.4	79.3	70.3	76.8	62.7	60.7	45.3	38.5	29.8	0.0	0.0	45.5	29.1	74.7	55.7	63.1	53.2
ABMDRNet [10]	ResNet-50	CVPR'21	94.3	84.8	90.0	69.6	75.7	60.3	64.0	45.1	44.1	33.1	31.0	5.1	61.7	47.4	66.2	50.0	69.5	54.8
CENet [31]	ResNet	RAL'23	92.0	85.8	78.9	70.0	74.9	61.4	64.8	46.8	39.8	29.3	65.7	8.7	54.1	47.8	77.1	56.9	71.8	56.1
CACFNet [17]	ConvNeXt-B	TIV'23	95.9	89.2	**93.6**	69.5	82.0	63.3	74.0	46.6	49.0	32.4	45.8	7.9	69.8	54.9	82.1	58.3	**76.7**	57.8
MMSMCNet [18]	MiT-B3	TCSVT'23	96.2	89.2	93.2	69.1	83.4	63.5	74.4	46.4	**56.6**	**41.9**	26.9	8.8	70.2	48.8	77.5	57.6	75.2	58.1
EAEFNet [19]	ResNet-152	RAL'23	95.4	87.6	85.2	72.6	79.9	63.8	70.6	48.6	47.9	35.0	62.8	14.2	62.7	52.4	71.9	58.3	75.1	58.9
CMX [13]	MiT-B2	TITS'23	92.2	89.4	81.3	74.8	73.4	64.7	63.5	47.3	38.8	30.1	36.3	8.1	53.3	52.4	67.7	59.4	67.3	58.2
CMX [13]	MiT-B4	TITS'23	-	90.1	-	75.2	-	64.5	-	50.2	-	35.3	-	8.5	-	54.2	-	60.6	-	59.7
CMNeXt [24]	MiT-B4	CVPR'23	94.4	**90.2**	83.9	74.2	77.3	63.8	55.7	45.4	47.5	38.1	32.1	13.4	55.8	51.8	63.8	58.6	67.8	59.3
CAINet [20]	MobileNet-V2	TMM'24	93.0	88.5	74.6	66.3	**85.2**	**68.7**	65.9	**55.4**	34.7	31.5	65.6	9.0	55.6	48.9	**85.0**	60.7	73.2	58.6
EGFNet [21]	ConvNeXt	TITS'24	**96.5**	89.8	92.1	71.6	84.8	63.9	**76.1**	46.7	44.6	31.3	38.7	6.7	**71.1**	52.0	78.1	57.4	75.6	57.5
Ours-dice	MiT-B2		93.1	88.8	89.2	74.5	76.4	64.5	69.3	47.8	52.0	36.6	25.6	**22.0**	64.7	53.7	70.0	60.9	71.0	**60.8**
Ours-con	MiT-B2		93.7	89.1	88.5	74.7	79.4	65.6	70.6	46.3	44.2	37.6	62.2	11.1	62.7	54.5	67.1	**62.0**	74.2	59.9
Ours-con-acc	MiT-B2		94.0	88.9	86.7	**75.4**	76.6	63.4	67.0	47.0	46.7	39.0	49.3	12.6	61.3	**55.7**	69.6	59.9	72.3	60.0

TABLE V
THE COMPARATIVE PER-CLASS RESULTS BASED ON RTFNET [9]. WE USE ACC (%) AND IOU (%) FOR EACH CLASS AND THE MACC (%) AND MIOU (%) FOR ALL THE CLASSES. WE USE RESNET-50 AND RESNET-152 AS BACKBONE TO COMPARE RTFNET AND OUR METHOD. THE RESULTS DEMONSTRATE THE SCALABILITY OF OUR METHOD, WITH THE BEST RESULTS OF EACH BACKBONE ARE HIGHLIGHTED IN **BOLD**.

Method	Backbone	Car		Person		Bike		Curve		Car Stop		Guardrail		Color Cone		Bump		mAcc	mIoU
		Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
RTFNet [9]	ResNet-50	**91.3**	86.3	78.2	67.8	71.5	58.2	59.8	43.7	32.1	**24.3**	13.4	3.6	40.4	26.0	**73.5**	**57.2**	62.2	51.7
Ours-con	ResNet-50	**91.3**	**87.5**	**85.8**	**71.1**	**73.9**	**60.6**	**61.7**	**43.8**	**35.0**	22.5	**43.3**	**3.8**	**45.4**	**38.4**	70.0	54.5	**67.3**	**53.3**
RTFNet [9]	ResNet-152	**93.0**	**87.4**	79.3	70.3	**76.8**	**62.7**	60.7	45.3	38.5	**29.8**	0.0	0.0	45.5	29.1	**74.7**	**55.7**	63.1	53.2
Ours-con	ResNet-152	91.5	86.5	**83.3**	**71.3**	74.5	61.7	**67.5**	**47.0**	**43.4**	28.4	**14.1**	**2.4**	**45.7**	**41.2**	73.2	50.5	**65.8**	**54.1**

loss and our proposed consistency loss on Img_2. Among them, the Dice loss focuses on improving the segmentation accuracy of the current frame, while the consistency loss focuses on improving segmentation consistency. The total loss is calculated as:

$$\mathcal{L}_{total} = \mathcal{L}_{seg1} + \alpha\mathcal{L}_{dice} + \mathcal{L}_{seg2} + \beta\mathcal{L}_{con}, \tag{4}$$

where α and β denote the importance of Dice loss and consistency loss, respectively. We set α and β to 0.5, 1.0, and 2.0 in turn, to investigate their influences on the segmentation results.

As shown in Tab. III, variant (E) obtains the highest mIoU, TC, and CA, indicating that the method achieves the best segmentation accuracy and consistency when $\alpha = 1.0$ and $\beta = 1.0$. Comparing variants (G) and (I), we find that variant (I) has much higher TC and CA but much lower mIoU than variant (G). This means that variant (I) achieves better consistency but worse accuracy, indicating that a higher β can enhance segmentation consistency while suppressing accuracy. Meanwhile, variant (G) achieves higher accuracy but at the cost of consistency, indicating that overemphasizing segmentation accuracy may hinder consistency, which is consistent with the conclusion from the previous results. The comparison between variants (A) and (C), and variants (D) and (F) leads to the same conclusion. This also demonstrates that balancing segmentation accuracy with consistency is a challenging problem and worth further investigating. In summary, when $\alpha = 1$ and $\beta = 1$, our method achieves the highest mIoU, TC, and CA at the same time, which indicates that our method and our loss function effectively balance segmentation accuracy and consistency.

D. Comparative Experiments

We adopt CMX with MiT-B2 as the segmentation network for our framework. Based on different loss function strategies, we select variants (G), (H) and (I) in Tab. I for comparison, naming them Ours-dice, Ours-con and Ours-con-acc. They all use MiT-B2 as the backbone. We compare our method with RTFNet [9], ABMDRNet [10], CENet [31], CACFNet [17], MMSMCNet [18], EAEFNet [19], CMX [13], CMNeXt [24], CAINet [20], and EGFNet [21]. CMX utilizes both MiT-B2 and MiT-B4 as backbones, while CMNeXt uses MiT-B4 as its backbone. Since CMX with MiT-B2 has not reported its per-class Acc, and CMNeXt with MiT-B4 has not reported its per-class Acc and IoU, we get the missing results by testing the networks with their pre-trained weights. However, CMX with MiT-B4 does not provide any pre-trained weights, so we could not compare its Acc with the other methods.

As shown in Tab. IV, our proposed methods achieve the highest mIoU among the state-of-the-art methods. Specifically, compared with CMX using the same MiT-B2 backbone, our methods achieve a 1.7% to 2.6% higher mIoU. Compared with CMNeXt that uses the larger backbone MiT-B4, our methods also increase mIoU by 0.2% to 1.1%. Moreover, the mAcc of Ours-con with MiT-B2 is 6.9% higher than that of CMX with MiT-B2. Despite achieving the highest mAcc, CACFNet [17] suffers from a low mIoU, which implies a high rate of false positives. Therefore, we can conclude that our method not only improves the segmentation consistency but also achieve the best overall accuracy compared to other methods.

To demonstrate the scalability of our method to other RGB-T semantic segmentation networks, we choose the well-known CNN-based RTFNet [9] as our segmentation network and keep the parameter quantities unchanged. Tab. V compares the

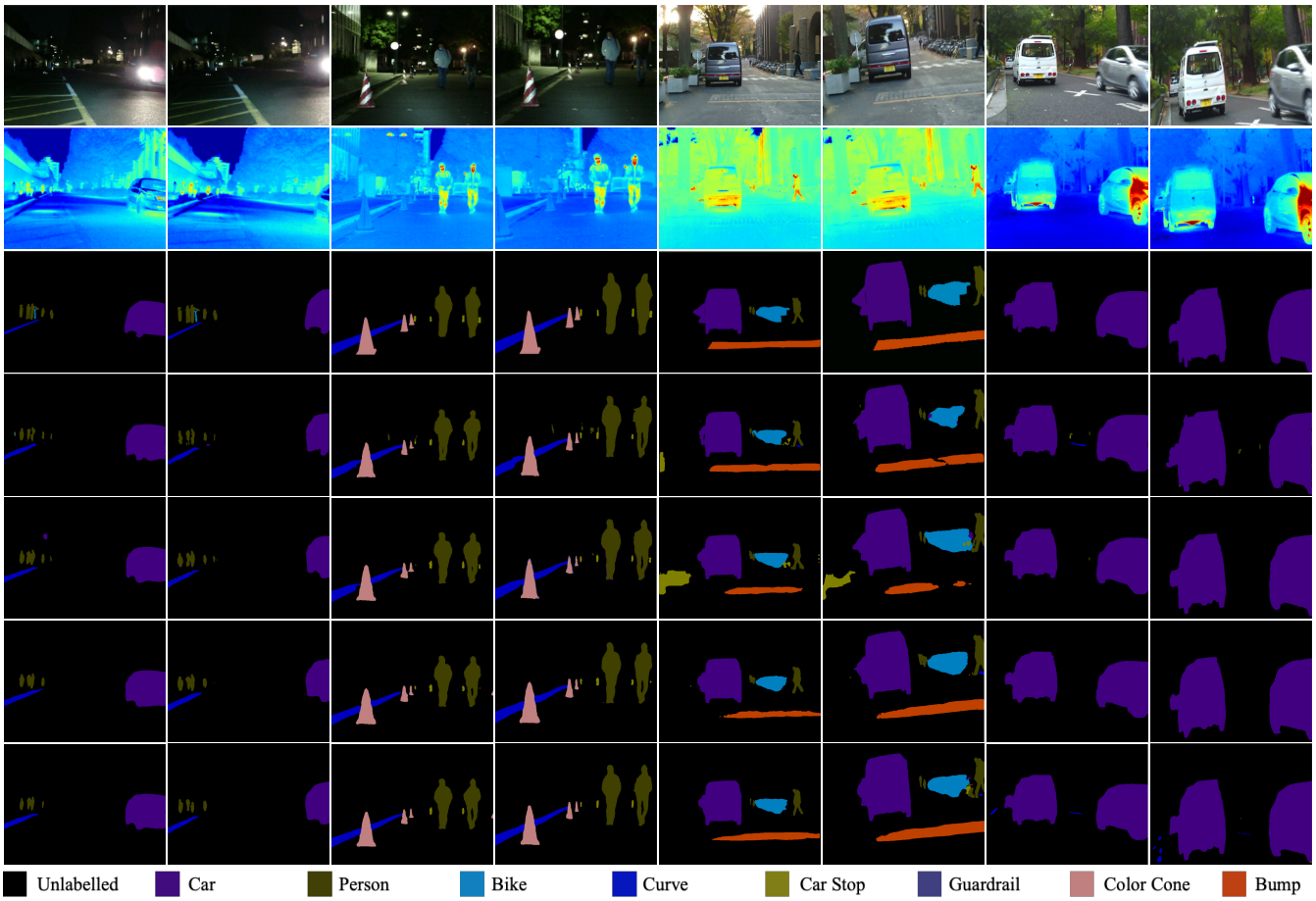


Fig. 6. Qualitative demonstration for RGB-T semantic segmentation on the MFNet [8] dataset. The rows from top to bottom are RGB images, thermal images, ground truth, `CMX` results, `Ours-dice` results, `Ours-con` results, and `Ours-con-acc` results. All of the variants use MiT-B2 as backbones. The 1st and 2nd columns are the 1st set of images, the 3rd and 4th columns are the 2nd set of images, the 5th and 6th columns are the 3rd set of images, and the 7th and 8th columns are the 4th set of images. Among them, the first two sets and the last two sets are the images of nighttime and daytime, respectively.

segmentation results of the original RTFNet and `Ours-con` based on different backbones. With ResNet-50 as the backbone, `Ours-con` outperforms RTFNet in segmentation accuracy (Acc and IoU) for all classes except car stop and bump, and increases mAcc and mIoU by 5.1% and 1.6%, respectively. With ResNet-152 as the backbone, `Ours-con` significantly increases the Acc of guardrail from 0.0% to 14.1%, and the IoU of color cone from 29.1% to 41.2%, indicating that our method can significantly enhance the segmentation accuracy for challenging objects. `Ours-con` with ResNet-152 also achieves better segmentation accuracy that increases mAcc and mIoU by 2.7% and 0.9%, respectively.

E. Qualitative Demonstrations

Fig. 6 qualitatively demonstrates sample segmentation results. In each set of images, the left column is the original current frame from the MFNet [8] dataset, and the right column is the synthesized frame generated by our method. The segmentation consistency can be seen by comparing the consecutive frames. We can see that our three methods generally outperform `CMX` with MiT-B2 in terms of segmentation accuracy, especially for small objects such as car stops (refer to the 3rd and 4th columns). We can also find that `Ours-con` achieves the most consistent results. Moreover, the results of all three of our methods are more consistent than those of `CMX`. Specifically, the segmentation results of `CMX` show significant inconsistencies for all sets of images. Similarly, the segmentation results of `Ours-dice` for 1st, 2nd, and 3rd sets are also inconsistent. The same is observed for `Ours-con-acc` in the 2nd, 3rd, and 4th sets. Although `Ours-con` presents the lowest mIoU among our methods, it shows much better consistency than the other methods. This result is consistent with the findings from the ablation experiments presented in Tab. I.

`Ours-con`, which uses the proposed consistency loss, achieves the best segmentation accuracy and has much better consistency than the other methods. So, in real applications, `Ours-con` could be preferable to `Ours-dice`, which sacrifices consistency for accuracy.

V. Conclusions and Future Work

This letter presents the viewpoint that consistency should be valued in RGB-T semantic segmentation in addition to accuracy. To this end, we proposed a temporal-consistent framework for consistent and accurate RGB-T semantic segmentation. The proposed framework includes a VVIG module, which can synthesize a virtual frame at the next moment. Moreover, the proposed consistency loss improves segmentation consistency

without compromising accuracy. The proposed metric is able to evaluate segmentation results in terms of both accuracy and consistency. The experimental results show that our method outperforms the state-of-the-art networks. In future work, we will investigate advanced techniques to better leverage the spatial consistency and temporal continuity of RGB-T information across consecutive frames. Moreover, we will extend our methods to other scenarios that require temporal consistency.

APPENDIX

Assume that a 3-D object point P is captured in two consecutive frames (Img_1 and Img_2), and the pixels for P on the two images are p_1 and p_2. From the pinhole camera model, we have: $p_1 \simeq KP, p_2 \simeq K(RP + t)$, where K is the camera intrinsic matrix, R and t are respectively the rotation matrix and translation vector between the consecutive frames, $\simeq$ denotes equivalence up to a scale.

Assume that the camera motion includes only rotations (i.e., $t = 0$), we have $p_2 \simeq KRP$. Since $P \simeq K^{-1}p_1$, we have $p_2 \simeq KRK^{-1}p_1$. Since $R = R_z(\alpha)R_y(\beta)R_x(\gamma)$, we have $p_2 \simeq KR_z(\alpha)R_y(\beta)R_x(\gamma)K^{-1}p_1$, where α, γ, β are respectively the Euler angles about the $\hat{Z}$, $\hat{Y}$, $\hat{X}$ axes. Since $p_2 \simeq M_{\mathrm{VVT}}p_1$, we have $M_{\mathrm{VVT}} = KR_z(\alpha)R_y(\beta)R_x(\gamma)K^{-1}$.

Comparing M_{VVT} to the homography matrix $H = K(R - \frac{tn^T}{d})K^{-1}$, where n^T and d are respectively the normal vector and the distance to the plane from the origin point, we can find that $M_{\mathrm{VVT}} = H$ if there is no translational movement (i.e., $t = 0$). So, we use the homography matrix as the VVT matrix in this work.

REFERENCES

[1] W. Ma, S. Huang, and Y. Sun, "Triplet-graph: Global metric localization based on semantic triplet graph for autonomous vehicles," *IEEE Robot. Autom. Lett.*, pp. 1–8, 2024.

[2] H. Xu, H. Liu, S. Huang, and Y. Sun, "C2l-pr: Cross-modal camera-to-lidar place recognition via modality alignment and orientation voting," *IEEE Trans. Intell. Veh.*, pp. 1–17, 2024.

[3] H. Xu, H. Liu, S. Meng, and Y. Sun, "A novel place recognition network using visual sequences and lidar point clouds for autonomous vehicles," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2023, pp. 2862–2867.

[4] Y. Feng and Y. Sun, "Polarpoint-bev: Bird-eye-view perception in polar points for explainable end-to-end autonomous driving," *IEEE Trans. Intell. Veh.*, pp. 1–11, 2024.

[5] S. Teng, X. Hu, P. Deng, B. Li, Y. Li, Y. Ai, D. Yang, L. Li, Z. Xuanyuan, F. Zhu, and L. Chen, "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Trans. Intell. Veh.*, vol. 8, no. 6, pp. 3692–3711, Jun. 2023.

[6] Y. Feng, W. Hua, and Y. Sun, "Nle-dm: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 9780–9791, 2023.

[7] H. Li and Y. Sun, "Igfnet: Illumination-guided fusion network for semantic scene understanding using rgb-thermal images," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2023, pp. 1–6.

[8] 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 *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2017, pp. 5108–5115.

[9] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, 2019.

[10] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, "Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2633–2642.

[11] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 3, pp. 1000–1011, 2021.

[12] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12 077–12 090, 2021.

[13] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers," *IEEE Trans. Intell. Transp. Syst.*, 2023.

[14] U. Shin, K. Lee, I. S. Kweon, and J. Oh, "Complementary random masking for rgb-thermal semantic segmentation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2024, pp. 11 110–11 117.

[15] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 686–695.

[16] D. Nilsson and C. Sminchisescu, "Semantic video segmentation by gated recurrent flow propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6819–6828.

[17] W. Zhou, S. Dong, M. Fang, and L. Yu, "Cacfnet: Cross-modal attention cascaded fusion network for rgb-t urban scene parsing," *IEEE Trans. Intell. Veh.*, 2023.

[18] W. Zhou, H. Zhang, W. Yan, and W. Lin, "Mmsmcnet: Modal memory sharing and morphological complementary networks for rgb-t urban scene semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[19] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng, and T. L. Lam, "Explicit attention-enhanced fusion for rgb-thermal perception tasks," *IEEE Robot. Autom. Lett.*, 2023.

[20] Y. Lv, Z. Liu, and G. Li, "Context-aware interaction network for rgb-t semantic segmentation," *IEEE Trans. Multimedia*, 2024.

[21] S. Dong, W. Zhou, C. Xu, and W. Yan, "Egfnet: Edge-aware guidance fusion network for rgb–thermal urban scene parsing," *IEEE Trans. Intell. Transp. Syst.*, 2023.

[22] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2021, pp. 10 347–10 357.

[23] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[24] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1136–1147.

[25] Y. Zhang, S. Borse, H. Cai, and F. Porikli, "Auxadapt: Stable and efficient test-time adaptation for temporally consistent video semantic segmentation," in *Proc. IEEE Win. App. Comput. Vis.*, 2022, pp. 2339–2348.

[26] Y. Liu, C. Shen, C. Yu, and J. Wang, "Efficient semantic video segmentation with per-frame inference," in *Proc. European Conf. Comput. Vis.* Springer, 2020, pp. 352–368.

[27] H. Park, A. Yessenbayev, T. Singhal, N. K. Adhikari, Y. Zhang, S. M. Borse, H. Cai, N. P. Pandey, F. Yin, F. Mayer, *et al.*, "Real-time, accurate, and consistent video semantic segmentation via unsupervised adaptation and cross-unit deployment on mobile device," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21 431–21 438.

[28] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vis.* IEEE, 2016, pp. 565–571.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, 2009, pp. 248–255.

[30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.

[31] Z. Feng, Y. Guo, and Y. Sun, "Cekd: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images," *IEEE Robot. Autom. Lett.*, vol. 8, no. 4, pp. 2205–2212, 2023.