

# CAFuser: Condition-Aware Multimodal Fusion for Robust Semantic Perception of Driving Scenes

Tim Brödermann<sup>1</sup>, Christos Sakaridis<sup>1</sup>, Yuqian Fu<sup>1,2</sup>, and Luc Van Gool<sup>1,2</sup>

**Abstract**—Leveraging multiple sensors is crucial for robust semantic perception in autonomous driving, as each sensor type has complementary strengths and weaknesses. However, existing sensor fusion methods often treat sensors uniformly across all conditions, leading to suboptimal performance. By contrast, we propose a novel, *condition-aware multimodal* fusion approach for robust semantic perception of driving scenes. Our method, CAFuser, uses an RGB camera input to classify environmental conditions and generate a *Condition Token* that guides the fusion of multiple sensor modalities. We further newly introduce modality-specific feature adapters to align diverse sensor inputs into a shared latent space, enabling efficient integration with a single and shared pre-trained backbone. By dynamically adapting sensor fusion based on the actual condition, our model significantly improves robustness and accuracy, especially in adverse-condition scenarios. CAFuser ranks first on the public MUSES benchmarks, achieving 59.7 PQ for multimodal panoptic and 78.2 mIoU for semantic segmentation, and also sets the new state of the art on DeLiVER. The source code is publicly available at: <https://github.com/timbroad/CAFuser>.

**Index Terms**—Sensor fusion, semantic scene understanding, computer vision for transportation, deep learning for visual perception, multimodal semantic perception

## I. INTRODUCTION

CURRENT perception pipelines for automated driving systems yield excellent results under normal, clear-weather conditions, but still struggle when they encounter adverse conditions. This prevents achieving the ultimate Level-5 driving automation, which requires a reliable perception system with an unlimited operational design domain (ODD). A major associated challenge is the accurate pixel-level semantic parsing of driving scenes, as experimental evidence [1] suggests that such high-level parsing is beneficial for the downstream driving tasks of prediction and planning.

Because of the aforementioned universal ODD requirement, using a single type of sensor for dense semantic perception of driving scenes is a fragile choice. More specifically, the sensitivity patterns of different sensors across environmental conditions differ drastically. For instance, while standard, RGB frame-based cameras have excellent spatial resolution, their

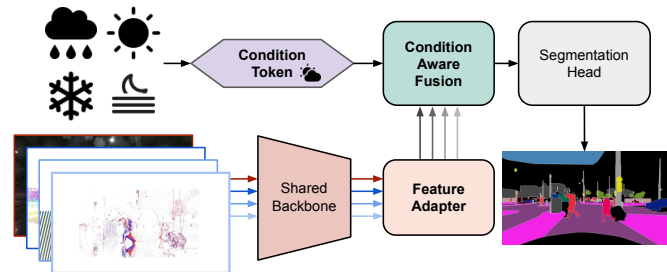


Fig. 1. CAFuser overview. We encode the weather and lighting conditions in a Condition Token and guide the condition-aware fusion with it.

measurements degrade severely in low illumination and adverse weather. Lidars and event cameras are by contrast much more robust to ambient lighting, but they are also strongly affected by weather particles such as raindrops or snowflakes. Radars handle adverse weather excellently but offer a far more limited spatial resolution. Thus, utilizing inputs from all sensor modalities of a vehicle’s suite in a multimodal fusion framework shows much more promise for reliable semantic perception, especially in adverse ODDs such as nighttime, fog, rain, or snowfall.

We observe that despite the decreasing cost of the aforementioned types of sensors and their growing adoption in autonomous vehicles, little attention has been paid to leveraging their complementary strengths in a *condition-aware* manner. That is, most current multimodal fusion methods fuse sensors *uniformly* across all environmental conditions. However, as the reliability of each sensor depends strongly on these conditions, fusing all sensors in a condition-agnostic fashion can generally lead to suboptimal performance. Thus, we propose a multimodal condition-aware fusion (CAF) module for robust semantic perception of driving scenes. By explicitly representing environmental conditions and adjusting the sensor fusion algorithm to these representations, we aim to arrive at an adaptive, condition-aware sensor fusion model, which knows its ODD and has learned which sensors are more informative and reliable for perception in that ODD.

Our network uses the RGB camera input to generate a *Condition Token* (CT), which guides the fusion of multiple sensors, ensuring that all sensors interact optimally. Learning this token is supervised with a verbo-visual contrastive loss utilizing text prompts à la CLIP [2], so as to ensure that the CT embeddings are aligned with the abstract language-based descriptions of environmental conditions, which can be provided as frame-level annotations of the multimodal input data [3]. By training the system end-to-end, our model learns to dynamically adapt to actual conditions, ensuring that

© 2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. This work was supported by the ETH Future Computing Laboratory (EFCL), financed by a donation from Huawei Technologies. (Corresponding authors: Tim Brödermann; Yuqian Fu.)

<sup>1</sup>Tim Brödermann, Christos Sakaridis, Yuqian Fu, and Luc Van Gool are with Computer Vision Laboratory, ETH Zurich, 8057 Zurich, Switzerland {timbr, csakarid, vangool}@vision.ee.ethz.ch.

<sup>2</sup>Yuqian Fu and Luc Van Gool are with INSAIT, Sofia University St. Kliment Ohridski, Bulgaria, {yuqian.fu, luc.vangool}@insait.ai.

sensor fusion is optimized for each condition, leading to more accurate semantic parsing.

Additionally, most existing fusion approaches rely on separate encoders for each sensor modality, leading to high computational complexity and requiring separate training pipelines for each modality. However, in real-time automated driving systems, efficiency is of paramount importance. In this regard, recent work [4] has demonstrated the effectiveness of large-scale pre-trained models even for non-RGB-camera modalities, showing the feasibility of using a shared backbone across multiple sensor modalities. This motivates us to introduce a *single network backbone* for extracting features from different sensor modalities, while still preserving the unique information from each modality.

Thus, our novel condition-aware multimodal fusion network design comprises a shared backbone for all modalities, combined with individual lightweight *feature adapters* [5] for each modality. We project all sensor inputs onto the image plane, as in prior multimodal segmentation approaches [6], [7], [8], in order to make them all compatible with the backbone they are fed to. Beyond being efficient, this design has two additional merits: 1) using the same backbone naturally allows non-RGB modalities such as lidar, radar, and event camera to get mapped to an RGB-compatible feature space, and 2) the feature adapters enable extraction of modality-specific information, providing complementary features to the RGB modality. Our experiments show a substantial reduction in model parameters (by 54%) via this design, while not sacrificing performance.

Extensive experiments show that our method, CAFuser (Condition-Aware Fuser), effectively learns to weigh different modalities in the fusion according to the present condition. CAFuser sets the new state of the art on DeLiVER [7] and ranks first on the public MUSES [3] benchmarks for both multimodal panoptic and semantic segmentation.

Our main contributions can be summarized as:

- **Condition-aware fusion:** We propose a *Condition Token* that guides our model to adaptively fuse based on the present environmental conditions, significantly improving robustness in adverse scenarios.
- **Efficient architecture:** We newly utilize feature adapters in a shared-backbone sensor fusion architecture. Each sensor’s features are aligned in a shared latent space, allowing efficient integration with the pre-trained backbone and reducing model size significantly.
- **Modular and scalable design:** Our architecture allows flexible and efficient addition of diverse sensor modalities, being adaptable to various sensor setups.
- **SOTA Performance:** We extensively ablate our method and demonstrate SOTA performance in multimodal panoptic and semantic segmentation.

## II. RELATED WORK

**Semantic perception** involves understanding the environment in a scene, encompassing tasks like semantic, instance, and panoptic segmentation. Semantic segmentation [9] focuses on classifying each pixel into a predefined category, while instance segmentation distinguishes between individual instances of objects within “things” (e.g., cars, pedestrians)

categories. Panoptic segmentation [10], [11] unifies semantic and instance segmentation by predicting both “stuff” (e.g., sky, road) and “things” in an image. These tasks are crucial for autonomous vehicles and robotics in general, as they provide the detailed environmental understanding necessary for downstream tasks such as path planning, obstacle avoidance, and applications like domestic service robotics [12]. Recent transformer-based architectures like MaskFormer [13] and Mask2Former [14] have advanced semantic perception. EfficientPS [15] improves efficiency without sacrificing accuracy, MaskDINO [16] predicts object masks simultaneously with bounding boxes, and OneFormer [17] achieves state-of-the-art results in instance, semantic, and panoptic segmentation with a single architecture and model. In contrast to these RGB-only methods, we tackle multimodal semantic perception by leveraging diverse sensor modalities for robust scene understanding in autonomous driving. Utilizing the recent MUSES dataset [3], which includes an RGB camera, a lidar, a radar, and an event camera, we enhance the reliability and accuracy of semantic perception.

**Multimodal feature fusion** is crucial in autonomous driving, as different sensors provide complementary information. Early research focused on enhancing lidar-based 3D detection with RGB camera data [18]. Large-scale datasets such as KITTI [19] and nuScenes [20] propelled this research but lacked recordings under adverse weather conditions, which motivated subsequent synthetic [7] and real-world [21], [22], [23], [3] datasets focusing on challenging environments.

Fusion techniques evolved from fusing two specific modalities [24], [25] to RGB-X fusion with arbitrary modalities [26], [27]. Methods like HRFuser [6] and CMNeXt [7] introduced architectures capable of handling multiple arbitrary sensor inputs, employing modular designs and attention mechanisms. StitchFusion [28] integrated large-scale pre-trained models directly as encoders and feature fusers, using a multi-directional adapter module for cross-modal information transfer during encoding. Recent works such as SAMFusion [29] explore modality-specific multimodal fusion for 3D detection, and [30] examines modality-agnostic multimodal fusion for semantic segmentation by employing a shared backbone but lacking feature adapters, leading to over-reliance on two modalities and no performance gain when adding lidar or event data. GeminiFusion [8] combines intra-modal and inter-modal attention to dynamically integrate complementary information across modalities, representing the state of the art in multimodal semantic segmentation. MUSES [3] tackles multimodal panoptic segmentation by local-window cross-attention to merge features from multiple independent backbones.

While these works enhance perception in challenging environments, they generally fuse sensor modalities uniformly and lack explicit adaptation to environmental conditions such as fog or low light. In contrast, we perform condition-aware fusion in a shared latent space, allowing our model to adapt dynamically to environmental conditions, thereby improving the robustness of multimodal semantic perception.

**Condition-aware perception** incorporates knowledge of the environmental conditions to guide perception. [31] tackles 2D object detection by implicitly allocating higher weights to the

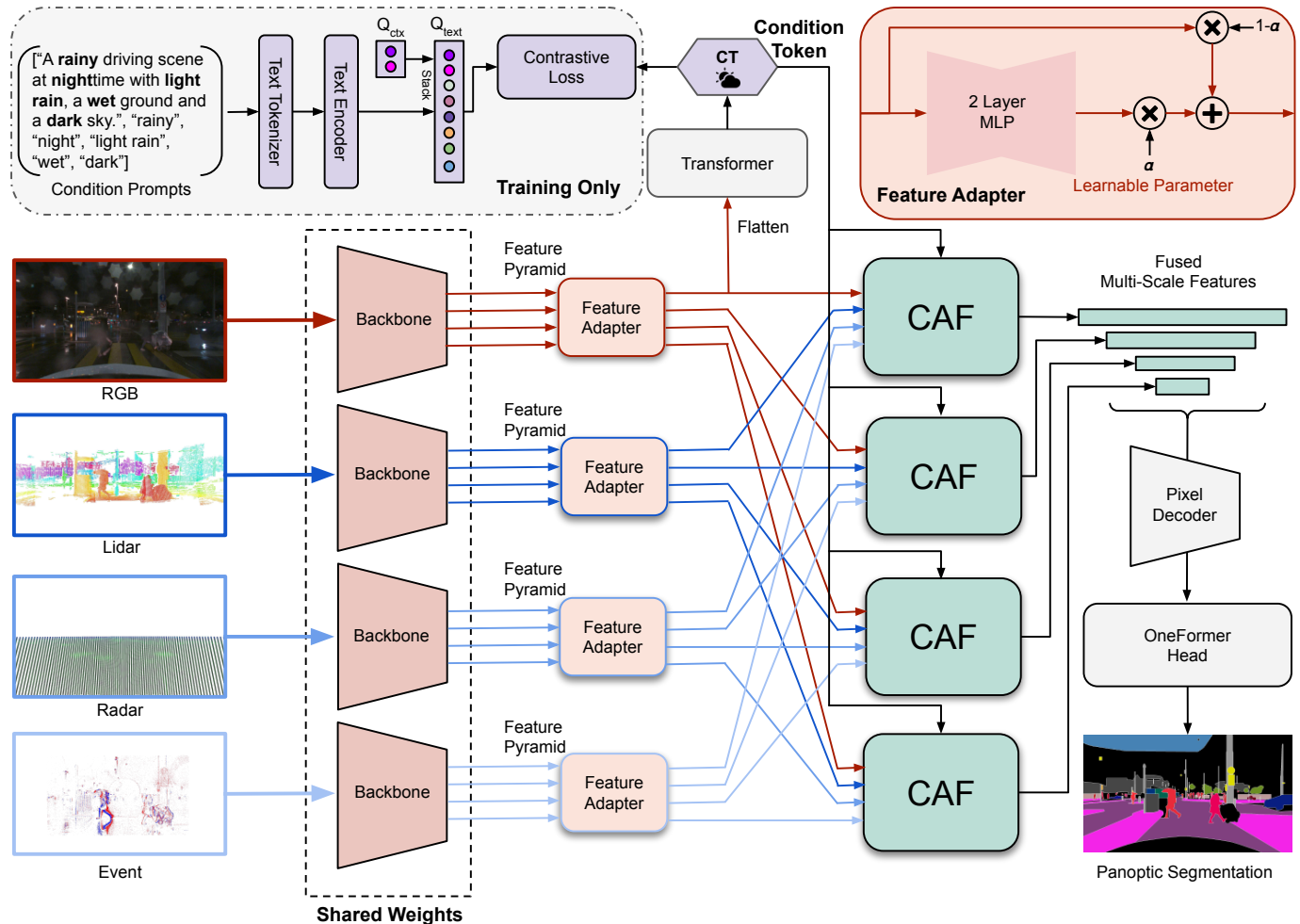


Fig. 2. **Our proposed CAFuser architecture** with RGB camera, lidar, radar, and event camera as input modalities. Each input is passed through the shared backbone and an individual feature adapter. The CT is generated from the highest-level RGB feature map, supervised with a verbo-visual contrastive loss to our encoded condition prompts ( $Q_{text}$ ), and used to guide the condition-aware fusion (CAF). The resulting fused multi-scale feature maps are then passed to the pixel decoder and the OneFormer [17] head to produce the prediction.

modality with better detection features at the late-stage fusion. CoLA [32] enhances salient object detection by leveraging pre-trained vision-language models with prompt learners to adjust for noisy or missing inputs. Knowledge distillation from vision-language models is used in [33] for modality-agnostic representations, without focusing on condition-aware predictions. By contrast, we explicitly encode detailed environmental conditions using a Condition Token supervised by verbal scene descriptions. This token dynamically guides our sensor fusion, improving the robustness of segmentation across diverse ODDs.

**Feature adapters** [5], [34], [35] offer a lightweight solution for integrating different sensor modalities into shared models. CLIP-Adapter [5] uses MLP adapters with residual connections to adapt features without overfitting. Modality-shared and modality-specific adapters were employed in [36] for RGB-thermal tracking. EventClip [4] aligns event data with CLIP features using feature adapters. While these methods focus on bimodal fusion, our network extends feature adapters to align diverse inputs from several modalities in a shared latent space, enabling flexible multimodal fusion.

### III. METHOD

We build on the recent OneFormer [17], using its head and segmentation framework as our starting point. In contrast to previous multimodal approaches, we introduce a single, shared backbone for all sensor modalities (see Fig. 2). By employing lightweight adapters for efficient feature transformation, this design significantly reduces model parameters while maintaining competitive performance. We initialize the backbone with ImageNet pre-training, which corresponds to the RGB camera modality. For pre-processing, we follow MUSES [3] and project each sensor’s data (e.g., lidar, radar) as a 3-channel image onto the RGB plane. Further, we normalize each modality over the entire dataset to ensure an input representation consistent with the RGB modality.

#### A. Multimodal Adapter

As the feature adapter, we employ a 2-layer MLP with a 4x reduction in hidden dimensions [5]. A learnable parameter  $\alpha$  controls the weighting of adapted and original features. Each modality and each feature map from the Swin backbone’s 4-level feature pyramid is adapted using an individual adapter at

each stage. For 4 modalities and 4 feature levels this results in 16 individual lightweight adapters. In Sec. IV-B we show that this setup allows us to reduce the parameters by 54% without any loss in performance.

### B. Condition-Aware Fusion

Since sensor reliability changes predictably based on environmental conditions, we introduce a condition-aware fusion (CAF) mechanism that dynamically adapts sensor fusion in response to the current ODD. As labeled condition data cannot be assumed to be available at inference time, we generate a *Condition Token* from the RGB camera input and use it to modulate the fusion process. The RGB camera captures sufficient global environmental information to effectively represent scene conditions, avoiding the computational overhead of processing additional modalities.

**Condition Token (CT):** Our CT generation, as shown in Fig. 2, starts by flattening the highest-level RGB feature map and passing it through a Transformer with 2 encoder and 2 decoder layers. The resulting CT is directly supervised during training using a verbo-visual contrastive loss utilizing text prompts based on a detailed description of the environmental condition. For this, the MUSES dataset provides several key scene attributes, including *weather condition*, *precipitation type* and *level*, *ground condition*, *time of day*, and *sky condition*. We automatically create a *condition prompt* from these attributes by slightly adapting them to fit into a continuous sentence. For example, the sky condition of *Sunlight* becomes *a sunny sky*. We further combine the precipitation type and level into one precipitation text (e.g. *light rain*) and fill in empty condition labels from context: e.g. at nighttime the sky condition attribute is often missing and is filled in as *dark*. Using these attributes, we construct a rich, descriptive condition prompt for each scene using the following template:

$$\begin{aligned} & \text{A } \{\text{weather condition}\} \text{ driving scene at} \\ & \{\text{time of day}\} \text{time with } \{\text{precipitation text}\}, \quad (1) \\ & \text{a } \{\text{ground condition}\} \text{ ground and a } \{\text{sun level}\} \text{ sky.} \end{aligned}$$

For example, our condition prompt may be instantiated as “A *rainy* driving scene at *nighttime* with *light rain*, a *wet* ground and a *dark* sky.” This detailed prompt is combined with the individual scene attributes and guides the CT to capture the nuanced environmental context needed for robust fusion.

We follow [17] to generate text queries ( $Q_{text}$ ) from our encoded condition prompts. This step includes a tokenizer for the condition prompts and a pass through a 6-layer transformer text encoder [37] and stacking with four context tokens  $Q_{ctx}$  [17], [38]. For the resulting  $Q_{text}$  and the CT, we apply a  $CT - Q_{text}$  contrastive loss [17], [39], [37].

At inference, we dynamically generate the CT directly from the RGB input, removing any reliance on explicit condition classification. This approach allows the CT to regulate the fusion on-the-fly based on the current environmental context, without any condition labels. Using the CT, we explore two fusion strategies:

**Condition-Aware Addition (CAA) Fusion:** In this simpler approach, we implement a CT-guided weighted addition fu-

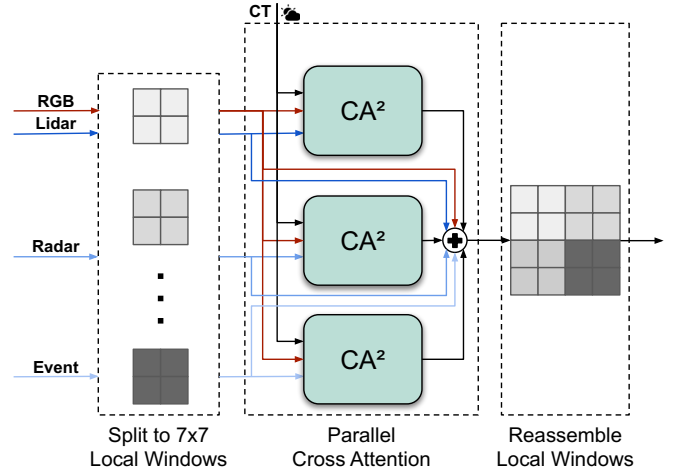


Fig. 3. **Condition-Aware Fusion (CAF)** for our  $CA^2$  variant. We apply multi-window cross-attention [6] by splitting each modality’s feature map into local windows, fusing all secondary modalities in parallel with the RGB features by using our proposed condition-aware cross-attention ( $CA^2$ ) module, and finally stitching the local windows back together.

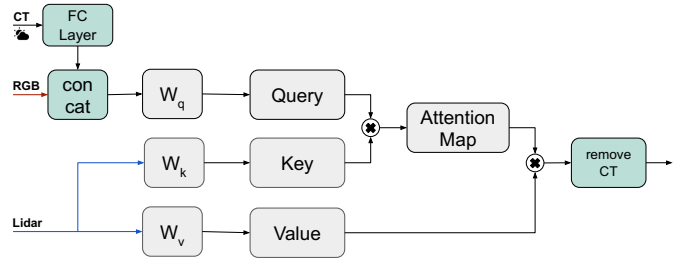


Fig. 4. **Condition-Aware Cross-Attention ( $CA^2$ )** applied to each local window, here illustrated for the case of RGB-lidar fusion. The Condition Token (CT) is passed through a fully connected layer to align with the feature dimension and concatenated with the RGB tokens to generate a condition-aware query for cross-attention. Afterwards, we remove the token corresponding to the CT to maintain the original spatial dimensions before reassembling the full feature map.

sion. We predict one weight per modality (totaling four modalities), ensuring that all weights sum to one. This is achieved by passing the flattened CT through a fully connected layer with four output dimensions, followed by a softmax function. After the feature adapters, we multiply each modality’s feature map by its corresponding predicted weight from the CT. We then sum the weighted feature maps per feature map level. The fused feature maps are then fed to the OneFormer head. This method already shows promising results by dynamically increasing the relevance of specific sensors based on different environmental conditions (cf. Sec. IV-B).

**Condition-Aware Cross-Attention ( $CA^2$ ) Fusion:** In our final model, CAFuser, we introduce  $CA^2$  Fusion, where the CT directly guides cross-attention to enhance modality fusion. As visualized in Fig. 3, we build upon the MWCA fusion block from [6], which utilizes a  $7 \times 7$  local-window cross-attention. In this block, we replace the standard cross-attention within each local window with our novel  $CA^2$  fusion, integrating the CT into the attention process. As depicted in Fig. 4, we first pass the CT through a fully connected layer to match the dimensionality of the appropriate feature map tokens. We then concatenate this adjusted CT with the 49 RGB tokens

from the local window, forming an enhanced, *condition-aware query* for cross-attention. This combined query captures both visual features and environmental conditions, which we use to compute the attention map applied to the value of the secondary modality (lidar, radar, etc.). After obtaining the attention output, we remove the token corresponding to the CT to maintain the original spatial dimensions before reassembling the full feature map. This condition-aware fusion approach leads to a significant performance boost, as evidenced by our experimental results in Sec. IV, where we observe notable improvements in segmentation accuracy under challenging conditions.

**Intuition:** In our approach, the RGB modality generates a query that interacts with the keys from secondary modalities via cross-attention. The intuition behind this design is that the RGB camera often contains high-quality visual information, but in challenging situations (e.g., a blurry region due to a raindrop on the lens), it might miss critical details. The cross-attention mechanism allows the RGB to “fill in the blanks” by looking up at the corresponding regions in the secondary modalities. Depending on the environmental condition (encoded in the CT) and in contrast to the CAA module, the model can focus more heavily on certain modalities for different regions within an image. For instance, in foggy conditions, the radar might be more reliable for distant objects, while in clear nighttime conditions, the lidar might be more useful. The combination of the RGB and the CT tokens enables the system to generate a dynamic attention map that adapts to the environment, ensuring robust performance across diverse conditions.

#### IV. EXPERIMENTS

We evaluate CAFuser on two multimodal driving datasets: MUSES [3] and DeLiVER [7]. **MUSES** contains 2500 real-world scenes with panoptic segmentation labels across eight conditions (day and night for rain, snow, fog, and clear weather). Each scene has RGB, lidar, radar, and event camera modalities, along with metadata describing the weather. We use the official SDK to project each modality onto the RGB plane. **DeLiVER** offers 7885 synthetic scenes for semantic segmentation in four adverse weather settings (cloudy, foggy, night, rainy), each with five corner-case artifacts (e.g. motion blur, lidar jitter). Each scene has RGB, projected lidar, depth, and event camera modalities. We generate a textual condition prompt similar to Sec. III-B with the template: “A synthetic [*condition*] driving scene with [*case*] artifacts.”.

CAFuser uses a Swin-T backbone [40] and follows the standard OneFormer training setup [17] with a batch size of 8. For MUSES, we train for 960 epochs with a 20% random drop of each modality [3] and for DeLiVER, we use 200k iterations, selecting the best checkpoint on the validation set.

##### A. Comparison to The State of The Art

We compare our model to state-of-the-art methods both in panoptic and semantic segmentation. Below, we summarize our results in comparison to other methods, showcasing the superior performance of CAFuser.

TABLE I  
COMPARISON OF PANOPTIC SEGMENTATION METHODS ON THE MUSES TEST SET IN PQ $\uparrow$ . \*: USES ONLY THE CAMERA MODALITY AS INPUT.

Method	Clear	Fog	Rain	Snow	Day	Night	All
Mask2Former [14]*	48.8	46.5	45.4	45.1	49.4	39.4	46.9
MaskDINO [16]*	54.1	46.2	46.23	48.54	51.9	42.7	49.4
OneFormer [17]*	58.3	53.7	53.4	53.8	57.6	47.8	55.2
HRFuser [6]	47.0	43.6	42.7	40.6	44.6	40.0	43.9
MUSES [3]	55.3	50.3	53.8	50.5	54.1	49.7	53.6
CAFuser-CAA	61.2	56.4	59.4	57.9	59.9	56.2	59.4
CAFuser-CA <sup>2</sup> (Ours)	61.4	57.5	59.6	57.2	59.5	57.3	59.7

TABLE II  
COMPARISON OF SEMANTIC SEGMENTATION METHODS ON THE MUSES TEST SET. C: RGB CAMERA, L: LIDAR, R: RADAR, E: EVENTS

Method	Modalities	Backbone	mIoU $\uparrow$
Mask2Former [14]	C	Swin-T	70.7
SegFormer [41]	C	MiT-B2	72.5
OneFormer [17]	C	Swin-T	72.8
CMNeXt [7]	CLRE	MiT-B2	72.4
GeminiFusion [8]	CLRE	MiT-B2	75.3
CAFuser-CAA	CLRE	Swin-T	78.5
CAFuser-CA <sup>2</sup> (Ours)	CLRE	Swin-T	78.2

In Table I, we compare our model with existing panoptic segmentation methods, both using RGB-camera-only inputs and multimodal inputs. For RGB-camera-only panoptic segmentation, we compare to strong baselines such as Mask2Former [14], MaskDINO [16], and OneFormer [17]. Among these methods, OneFormer achieves the highest performance, highlighting the strength of this architecture. In a multimodal setting, MUSES [3] achieves the highest PQ with 53.9%. However, our method, which introduces CAF, achieves a *new state-of-the-art* result with a PQ score of 59.7%. The performance gain is especially notable in the *night* split with +7.6% margin over the second best method, compared to +1.9% in the *day* split. Since each time-of-day split includes all weather conditions (clear, fog, rain, and snow), these findings underscore the importance of adapting sensor weights when the RGB modality is less reliable.

As multimodal panoptic segmentation is still an emerging field, we further benchmark our model against the state of the art in multimodal *semantic* segmentation on both MUSES in Table II and DeLiVER in Table III. Given the OneFormer head simultaneously solves panoptic and semantic segmentation, we obtain one CAFuser model solving both tasks on MUSES. The results show that our model outperforms top methods such as CMNeXt [7] and GeminiFusion [8]. While these multimodal methods do improve over the RGB-camera-only methods, our CAFuser model significantly outperforms all multimodal and RGB-only methods and sets the new SOTA. This validates the strength of CAF, which allows optimal sensor integration under diverse environmental conditions and thus superior performance both in panoptic and semantic segmentation.

##### B. Ablation Studies

We perform all our ablations on MUSES due to its diverse real-world conditions and high-quality panoptic annotations. **Feature Adapter:** We ablate the effect of our proposed feature adapter and CAF module in Tab. IV where we first create

TABLE III

COMPARISON OF SEMANTIC SEGMENTATION METHODS ON DELIVER. C: RGB CAMERA, D: DEPTH, E: EVENTS, L: LIDAR

Method	Modalities	Backbone	mIoU-val $\uparrow$	mIoU-test $\uparrow$
CMNeXt [7]	CLDE	MiT-B2	66.3	53.0
StitchFusion [28]	CLDE	MiT-B2	68.2	53.4
GeminiFusion [8]	CLDE	MiT-B2	66.9	54.5
CAFuser-CAA	CLDE	Swin-T	68.6	55.2
CAFuser-CA <sup>2</sup> (Ours)	CLDE	Swin-T	67.8	55.6

TABLE IV

ABLATION STUDY OF OUR PROPOSED MODULES ON MUSES. USING A SHARED BACKBONE REDUCES THE PARAMETERS BY 55% WHILE THE ADAPTER AND CAF MODULE INCREASE THE PQ TO A NEW SOTA.

Method	Shared Backb.	Adapter	CAF	Params	PQ $\uparrow$
1 OneFormer [17]	n/a	-	-	50.7M	55.7
2 OneFormer w/ MUSES	-	-	-	149.0M	59.3
3 CAFuser	$\checkmark$	-	-	66.4M	58.4
4 CAFuser	$\checkmark$	$\checkmark$	-	68.0M	59.3
5 CAFuser-CA <sup>2</sup> (Ours)	$\checkmark$	$\checkmark$	$\checkmark$	77.7M	59.7

a strong baseline (row 2) by training OneFormer with the 4 parallel backbones proposed by MUSES. Using a shared backbone (row 3) significantly reduces the number of parameters (-55%), but also results in a significant performance drop (-0.9% PQ). Adding our proposed feature adapter (row 4) gains back all the lost performance (+0.9% PQ) while still having less than half of the original parameters. Adding our CAF module (row 5) gains another 0.4% in PQ, surpassing the performance of the larger, baseline model.

**CAA Fusion:** We conduct an ablation study to evaluate the impact of our CAA fusion method. As shown in Table V, we compare four fusion strategies: mean, random weights, learned weights, and CAA Fusion, which dynamically predicts modality weights using the CT. For the basic method that takes a simple mean over all modalities, we can see a large improvement of +1% PQ when adding the adapter (row 2) to the no-adapter baseline, highlighting the benefit of adapting features before fusion. Random fusion, where weights are randomly assigned to each modality, underperformed with a PQ of 58.5%. Learning static weights for each modality, without considering environmental conditions, reaches a similar performance to the mean fusion. Finally, our proposed CAA Fusion, which dynamically adjusts weights based on the CT, outperforms all other strategies with a PQ of 59.4%, demonstrating the effectiveness of adapting the fusion process to environmental conditions.

**CA<sup>2</sup> Fusion:** In Tab. VI, we investigate the detailed design of our CA<sup>2</sup> fusion and how to best utilize the CT for CAF.

TABLE V

ABLATION STUDY USING ALL FOUR MODALITIES ON DIFFERENT WEIGHTED FUSION STRATEGIES FOR OUR CAA MODULE ON MUSES.

Adapter	CAF	Fusion Type	PQ $\uparrow$
1	$\times$	Mean	58.1
2	$\checkmark$	Mean	59.1
3	$\checkmark$	Random Weights	58.5
4	$\checkmark$	Learned Weights	59.1
5	$\checkmark$	CAA	59.4

TABLE VI

ABLATION STUDY ON CA<sup>2</sup> FUSION DESIGN ON MUSES. THE CONDITION TOKEN IS APPENDED TO THE RGB QUERIES (Q) OR THE SECONDARY MODALITIES' KEYS AND VALUES (K&V).

	Q	K&V	PQ $\uparrow$
1	$\times$	$\times$	59.3
2	$\times$	$\checkmark$	59.1
3 (Ours)	$\checkmark$	$\times$	59.7
4	$\checkmark$	$\checkmark$	59.6

TABLE VII

ABLATION STUDY ON THE GUIDANCE OF THE VERBO-VISUAL CONTRASTIVE LOSS BETWEEN  $Q_{text}$  AND THE CT ON MUSES.

	Condition Loss	PQ $\uparrow$
1	$\times$	59.3
2 (Ours)	$\checkmark$	59.7

Appending the CT to the secondary modalities' keys and values yields a small decrease in performance. Since the CT is derived from RGB features, mixing it with other modalities could introduce confusion to the fusion process.

In contrast, our proposed approach (row 3), where the CT is only appended to the RGB queries, achieves the highest PQ score of 59.7%. This design enables dynamic cross-attention tailored to environmental conditions, effectively guiding the fusion process and aligning with our intuition described in Sec. III-B and motivating our design choice.

**Condition Loss:** In Table VII, we assess the impact of our contrastive condition loss, by training an identical network, with and without applying the loss on the CT. As this loss is designed to ensure the CT effectively captures environmental conditions, we want to investigate if the model could learn this internally without additional supervision. Adding the condition loss improves the performance by +0.4% in PQ. This demonstrates that our contrastive loss is essential as an additional supervision to efficiently guide CAF.

**Condition Encoding:** In Table VIII, we ablate the use of different condition prompt encodings in the CT generation. Using a high-level condition description, such as simply classifying weather with a single token as e.g. *clear* or *foggy*, yields the same performance as without CAF (PQ 59.2%). However, our proposed detailed condition prompt construction, which cap-

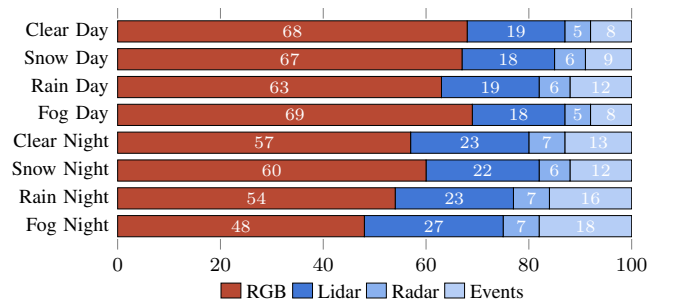


Fig. 5. Average Condition-Aware Addition (CAA) fusion weights in % on the MUSES test set across different weather conditions and times of day. This figure illustrates how the relative contributions of each sensor modality vary under various environmental conditions, highlighting the adaptability of the fusion mechanism to changing visibility and lighting scenarios.

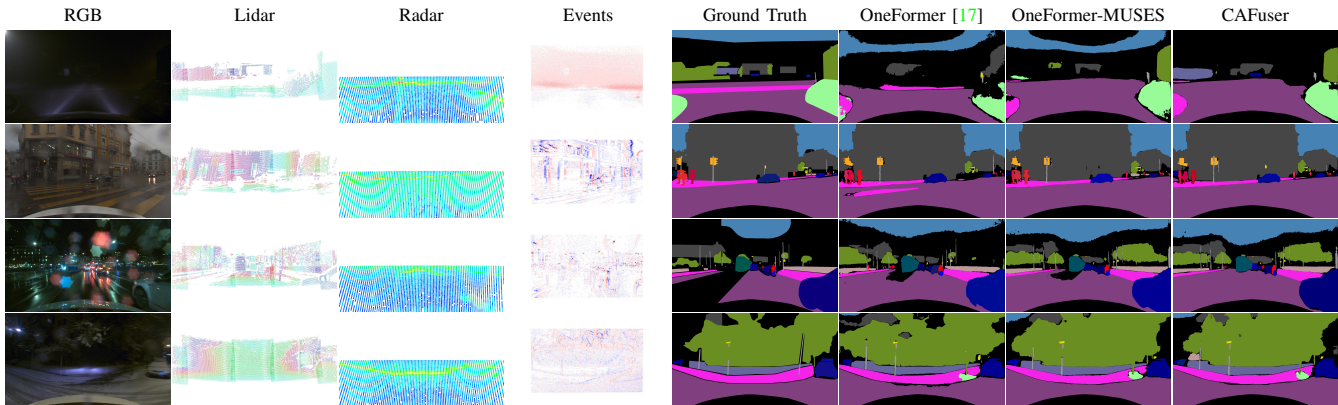


Fig. 6. Qualitative panoptic segmentation results on MUSES with visualization of the four input modalities. Best viewed on a screen at full zoom.

TABLE VIII

ABLATION STUDY ON THE CONDITION PROMPT ENCODINGS ON MUSES.

1: SIMPLE HIGH-LEVEL CONDITION PROMPT (E.G. ["CLEAR"]) 2: DETAILED CONDITION PROMPTS ACCORDING TO SEC. III-B.

	Text Tokens	PQ $\uparrow$
1	w/o detailed condition prompts	59.2
2 (Ours)	w/ detailed condition prompts	<b>59.7</b>

TABLE IX

ABLATION FOR CAFUSER WITH DIFFERENT INPUT MODALITIES ON MUSES. EACH ADDED MODALITY IMPROVES THE PQ.

	RGB	Lidar	Radar	Events	PQ $\uparrow$
1	✓	×	×	×	55.7
2	✓	✓	×	×	58.7 (+3.0)
3	✓	✓	✓	×	59.3 (+0.6)
4	✓	✓	✓	✓	<b>59.7 (+0.4)</b>

tures nuanced aspects of the environment (e.g., distinguishing between snow on the ground versus snow in the air), results in a significant improvement of +0.5% in PQ, achieving a PQ of 59.7%. This suggests that detailed condition prompts provide more fine-grained information for the fusion process, enabling more effective sensor integration based on the actual environmental context.

**Modalities:** Table IX shows the performance of our model with different combinations of input modalities. Starting with RGB-only input (55.7% PQ), we observe consistent performance gains with each added modality: +3.0% PQ for lidar, +0.6% PQ for radar, and an additional +0.4% PQ for event camera, reaching 59.7% PQ. These results demonstrate that our approach effectively leverages all modalities, with each additional sensor contributing to performance. The flexibility of our design allows to generalize well across various input combinations.

### C. CAA Weight Analysis

We also investigate how our CAA fusion mechanism dynamically adjusts sensor modality weights in response to changing environmental conditions in scenes that were not encountered during training. Fig. 5 illustrates that under clear

daytime conditions, the RGB modality dominates with a weight of 68%, capitalizing on its rich visual detail. In contrast, in foggy nighttime scenes, the RGB weight significantly decreases by -20% to 48%, while the weights of other modalities increase correspondingly by +8% for lidar, +2% for radar, and +10% for the events. This substantial shift indicates that our CAA module effectively adapts to challenging conditions by decreasing reliance on less reliable sensors and boosting the weight of more robust ones per case. These adjustments mirror the natural strengths of each sensor modality. Since RGB cameras struggle in low light and in certain adverse weather conditions, the CAA fusion places greater reliance on the active sensors (lidar and radar) and event cameras, which offer a high dynamic range and excel in low-light environments. These findings highlight that we successfully encode environmental conditions in our CT from RGB inputs alone and adjust the fusion mechanism to prioritize the most informative sensors, thereby making semantic perception in automated driving more robust.

### D. Qualitative Results

In Fig. 6, we compare qualitative results of CAFuser with competing methods. In the foggy nighttime scene (row 1), CAFuser successfully identifies the distant car in the center. In row 2, both multimodal approaches label the person on the right, despite a droplet obscuring the RGB input. In row 3, only CAFuser segments the more distant of the two riders waiting at the traffic light. In row 4, while all methods recognize the four vehicles on the sides, only CAFuser achieves good instance separation without unnecessary gaps.

## V. CONCLUSION

In this work, we introduced CAFuser, a novel condition-aware multimodal fusion framework for robust semantic perception in autonomous driving. By employing a shared backbone and modality-specific feature adapters, CAFuser efficiently aligns diverse sensor inputs into a common latent space while significantly reducing the model complexity. Our attention-based condition-aware fusion module dynamically adapts to environmental conditions guided by a Condition Token that is learned from the RGB input. This dynamic fusion

enhances robustness and accuracy under challenging weather scenarios. We demonstrated that our method outperforms competing approaches both in panoptic and semantic segmentation, setting the new state of the art on MUSES and DeLiVER. Extensive ablation studies demonstrate the effectiveness of our proposed modules. These advancements make CAFuser a promising solution for enhancing semantic perception in autonomous driving and robotics, particularly under adverse environmental conditions.

## REFERENCES

- [1] B. Zhou, P. Krähenbühl, and V. Koltun, “Does computer vision matter for action?” *Science Robotics*, vol. 4, no. 30, 2019. [1](#)
- [2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763. [1](#)
- [3] T. Brödermann, D. Brügemann, C. Sakaridis, K. Ta, O. Liagouris, J. Corkill, and L. Van Gool, “Muses: The multi-sensor semantic perception dataset for driving under uncertainty,” in *European Conference on Computer Vision (ECCV)*, 2024. [1](#), [2](#), [3](#), [5](#)
- [4] Z. Wu, X. Liu, and I. Gilitschenski, “Eventclip: Adapting clip for event-based object recognition,” *arXiv preprint arXiv:2306.06354*, 2023. [2](#), [3](#)
- [5] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, “Clip-adapter: Better vision-language models with feature adapters,” *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024. [2](#), [3](#)
- [6] T. Broedermann, C. Sakaridis, D. Dai, and L. Van Gool, “HRFuser: A multi-resolution sensor fusion architecture for 2D object detection,” in *ITSC*, 2023. [2](#), [4](#), [5](#)
- [7] 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 *CVPR*, 2023. [2](#), [5](#), [6](#)
- [8] D. Jia, J. Guo, K. Han, H. Wu, C. Zhang, C. Xu, and X. Chen, “Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer,” *arXiv preprint arXiv:2406.01210*, 2024. [2](#), [5](#), [6](#)
- [9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The Cityscapes dataset for semantic urban scene understanding,” in *CVPR*, 2016. [2](#)
- [10] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, “Panoptic segmentation,” in *CVPR*, 2019. [2](#)
- [11] O. Elharrouss, S. Al-Maadeed, N. Subramanian, N. Ottakath, N. Al-maadeed, and Y. Himeur, “Panoptic segmentation: A review,” *arXiv preprint arXiv:2111.10250*, 2021. [2](#)
- [12] J. V. Hurtado and A. Valada, “Semantic scene segmentation for robotics,” in *Deep learning for robot perception and cognition*. Elsevier, 2022, pp. 279–311. [2](#)
- [13] B. Cheng, A. Schwing, and A. Kirillov, “Per-pixel classification is not all you need for semantic segmentation,” in *NeurIPS*, 2021. [2](#)
- [14] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” in *CVPR*, 2022. [2](#), [5](#)
- [15] R. Mohan and A. Valada, “Efficientps: Efficient panoptic segmentation,” *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1551–1579, 2021. [2](#)
- [16] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, “Mask dino: Towards a unified transformer-based framework for object detection and segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3041–3050. [2](#), [5](#)
- [17] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, “OneFormer: One transformer to rule universal image segmentation,” in *CVPR*, 2023. [2](#), [3](#), [4](#), [5](#), [6](#), [7](#)
- [18] M. Liang, B. Yang, S. Wang, and R. Urtasun, “Deep continuous fusion for multi-sensor 3D object detection,” in *ECCV*, 2018. [2](#)
- [19] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite,” in *CVPR*, 2012. [2](#)
- [20] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuScenes: A multimodal dataset for autonomous driving,” in *CVPR*, 2020. [2](#)
- [21] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, “Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather,” in *CVPR*, 2020. [2](#)
- [22] M. Pitropov, D. E. Garcia, J. Rebello, M. Smart, C. Wang, K. Czarnicki, and S. Waslander, “Canadian adverse driving conditions dataset,” *The International Journal of Robotics Research*, vol. 40, no. 4-5, pp. 681–690, 2021. [2](#)
- [23] C. Sakaridis, D. Dai, and L. Van Gool, “ACDC: The Adverse Conditions Dataset with Correspondences for semantic driving scene understanding,” in *ICCV*, 2021. [2](#)
- [24] M. Pollach, F. Schiegg, and A. Knoll, “Low latency and low-level sensor fusion for automotive use-cases,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020. [2](#)
- [25] S. S. Chaturvedi, L. Zhang, and X. Yuan, “Pay” attention” to adverse weather: Weather-aware attention-based object detection,” in *ICPR*. IEEE, 2022. [2](#)
- [26] 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 Transactions on Intelligent Transportation Systems*, 2023. [2](#)
- [27] J. Huang, J. Li, N. Jia, Y. Sun, C. Liu, Q. Chen, and R. Fan, “Roadformer+: Delivering rgb-x scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion,” *IEEE Transactions on Intelligent Vehicles*, 2024. [2](#)
- [28] B. Li, D. Zhang, Z. Zhao, J. Gao, and X. Li, “Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation,” *arXiv preprint arXiv:2408.01343*, 2024. [2](#), [6](#)
- [29] E. Palladin, R. Dietze, P. Narayanan, M. Bijelic, and F. Heide, “SAM-Fusion: Sensor-adaptive multimodal fusion for 3D object detection in adverse weather,” in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2024. [2](#)
- [30] X. Zheng, Y. Lyu, J. Zhou, and L. Wang, “Centering the value of every modality: Towards efficient and resilient modality-agnostic semantic segmentation,” in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2024. [2](#)
- [31] S. S. Chaturvedi, L. Zhang, and X. Yuan, “Pay” attention” to adverse weather: Weather-aware attention-based object detection,” in *ICPR*. IEEE, 2022. [2](#)
- [32] S. Hao, C. Zhong, and H. Tang, “Cola: Conditional dropout and language-driven robust dual-modal salient object detection,” in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2024. [3](#)
- [33] X. Zheng, Y. Lyu, and L. Wang, “Learning modality-agnostic representation for semantic segmentation from any modalities,” in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2024. [3](#)
- [34] L. Yang, R.-Y. Zhang, Y. Wang, and X. Xie, “Mma: Multi-modal adapter for vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 826–23 837. [3](#)
- [35] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, “Vision transformer adapter for dense predictions,” *arXiv preprint arXiv:2205.08534*, 2022. [3](#)
- [36] A. Lu, C. Li, Y. Yan, J. Tang, and B. Luo, “Rgbt tracking via multi-adapter network with hierarchical divergence loss,” *IEEE Transactions on Image Processing*, vol. 30, pp. 5613–5625, 2021. [3](#)
- [37] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, “Groupvit: Semantic segmentation emerges from text supervision,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 134–18 144. [4](#)
- [38] A. Das, X. Hu, L. Jiang, and B. Schiele, “MTA-CLIP: Language-guided semantic segmentation with mask-text alignment,” in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2024. [4](#)
- [39] D. Brüggemann, C. Sakaridis, T. Brödermann, and L. Van Gool, “Contrastive model adaptation for cross-condition robustness in semantic segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 378–11 387. [4](#)
- [40] 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 *ICCV*, 2021. [5](#)
- [41] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” in *NeurIPS*, 2021. [5](#)