This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2024.3454971

CLFT: Camera-LiDAR Fusion Transformer for Semantic Segmentation in Autonomous Driving

Junyi Gu, Mauro Bellone, Tomáš Pivoňka, and Raivo Sell

***Abstract*—Critical research about camera-and-LiDAR-based semantic object segmentation for autonomous driving significantly benefited from the recent development of deep learning. Specifically, the vision transformer is the novel ground-breaker that successfully brought the multi-head-attention mechanism to computer vision applications. Therefore, we propose a vision-transformer-based network to carry out camera-LiDAR fusion for semantic segmentation applied to autonomous driving. Our proposal uses the novel progressive-assemble strategy of vision transformers on a double-direction network and then integrates the results in a cross-fusion strategy over the transformer decoder layers. Unlike other works in the literature, our camera-LiDAR fusion transformers have been evaluated in challenging conditions like rain and low illumination, showing robust performance. The paper reports the segmentation results over the vehicle and human classes in different modalities: camera-only, LiDAR-only, and camera-LiDAR fusion. We perform coherent controlled benchmark experiments of the camera-LiDAR fusion transformer (CLFT) against other networks that are also designed for semantic segmentation. The experiments aim to evaluate the performance of CLFT independently from two perspectives: multimodal sensor fusion and backbone architectures. The quantitative assessments show our CLFT networks yield an improvement of up to 10% for challenging dark-wet conditions when comparing with Fully-Convolutional-Neural-Network-based (FCN) camera-LiDAR fusion neural network. Contrasting to the network with transformer backbone but using single modality input, the all-around improvement is 5-10%.**

Our full code is available online for an interactive demonstration and application[1].

***Index Terms*—Camera-LiDAR fusion, Transformer, Semantic Segmentation, Autonomous driving.**

I. Introduction

Semantic segmentation of the surrounding environment is a challenging topic in autonomous driving and plays a critical role in various intelligent-vehicle-related research-tasks such as maneuvering, path planning [1] [2], and scene understanding [3]. The field of semantic segmentation has greatly advanced due to the evolution of deep neural networks, particularly Convolutional Neural Networks (CNN), along with the availability of open datasets. Early studies [4] took camera RGB images as input and tested them with datasets that had relatively monotonous scenarios [5]. In recent years, the blooming of perceptive sensor industries and strict safety requirements motivated semantic segmentation research related to different sensors and comprehensive scenarios. LiDAR sensors are involved the most in all kinds of research. Examples of the popular LiDAR-only methods include VoxNet [6], PointNet [7], and RotationNet [8]. However, multimodal sensor fusion is perceived as a promising technique to solve the problem of autonomous driving and has become the mainstream option for semantic segmentation [9].

As an applied research, the advancement of semantic segmentation is driven by the proposals of neural network backbones. One of the most popular neural networks recently proposed is the transformer [10], which implemented the multi-head attention mechanism [11] into the Natural Language Processing (NLP) application. The proposal of the Vision Transformer (ViT) [12] inspired researchers to explore its potential in environment perception for autonomous driving. In this work, we introduce the camera-LiDAR fusion transformer (CLFT). CLFT maintains the generic encoder-decoder architecture of a transformer-based network but uses a novel progressive-assemble strategy of vision transformers on a double-direction network. The results of the two network directions are then integrated using a cross-fusion strategy over the transformer decoder layers.

The CLFT aims to address the following issues that are challenging and less explored in the autonomous driving community.

(i) **Unbalanced sample distribution.** In real-traffic scenarios, dealing with an unbalanced sample distribution poses a significant challenge for autonomous vehicles. For instance, while vehicle lanes consistently have more cars than humans (primarily encountered at crossings or sidewalks), achieving precise perception of human entities remains paramount for the optimal functioning of any autonomous vehicle. Our previous camera-LiDAR FCN-based fusion model (CLFCN) [13] achieved more than 90% accuracy in vehicle classification. However, its accuracy in the human class is limited, reaching only 50%. Due to the under-representation of the human class in the dataset, CNNs face challenges in effectively learning knowledge during explicit down-sampling processes. In contrast, vision transformers maintain a consistent resolution for representations across all stages. Furthermore, their incorporation of a multi-head self-attention mechanism inherently provides an advantage in handling global context, making them more adept at addressing challenges associated with imbalanced class distributions.

(ii) **The consistency of multimodal input data formats.**

Corresponding Author: Mauro Bellone

J. Gu (junyi.gu@taltech.ee) and R. Sell(raivo.sell@taltech.ee) are with the Department of Mechanical and Industrial Engineering, Tallinn University of Technology, Estonia.

M. Bellone (mauro.bellone@taltech.ee) is with FinEst Centre for Smart Cities, Tallinn University of Technology, Estonia.

Tomáš Pivoňka (tomas.pivonka@cvut.cz) is with Czech Institute of Informatics, Robotics, and Cybernetics and Department of Cybernetics, Czech Technical University in Prague, Czech Republic.

[1] https://github.com/Claud1234/CLFT

LiDAR sensors have attracted broad interest from autonomous driving community and there are different strategies to process the LiDAR's point clouds data [14]. Unlike previous works in this field that integrate a voxel view of the LiDAR with the camera view [15] [16], our work uses the strategy to project the LiDAR point clouds along XY, YZ, and XZ plane views; thus, the camera and LiDAR inputs are amalgamated into a unified data representation for subsequent operations, encompassing feature extraction, assembly, and fusion. Although our CLFT models require the pre-processing of LiDAR point clouds such as calibration, filtering, and projection, we have verified that it is possible to carry out all these operations on the fly based on the current hardware specifications on autonomous vehicles [17] without significant overhead. Together with the inference time analysis in Section V, it is possible to claim the practical potential applicability of our models.

The niche of our work compared to other state-of-the-art transformer-based multimodal fusion techniques is detailed in Section II. The contribution of this work can be summarily outlined as follows:

- We introduce a new network architecture named CLFT, employing an innovative progressive-assemble strategy of vision transformers within a double-direction network.
- To the best of our knowledge [18] [19], CLFT is the first open-source transformer-based network that directly uses camera and LiDAR sensory input for object semantic segmentation tasks.
- We divide datasets based on illumination and weather conditions. This approach allows us to compare and highlight the robustness and efficacy of different models in challenging real-world situations.
- We prove the advancement and prospect of multimodal transformer-based models in the autonomous driving perception field, especially the segmentation of under-represented traffic objects.

The remainder of the paper is as follows. Section II reviews the state-of-the-art literature on camera-LiDAR deep fusion and transformer usage in autonomous driving. We analyze the gap in current research and explain how our work contributes to the field. Section III introduces the CLFT architecture details. Section IV presents the pre-processing and configurations of the dataset we used in this work. Section V reports the experiment results and discussion. Finally, a conclusion is conducted in Section VI.

II. Related work

Given the scope of this work, we revisit relevant literature on two aspects of semantic object segmentation for autonomous driving. The first part reviews the popular camera-LiDAR fusion-based deep learning proposals. The second part presents the recent usage of transformers in autonomous driving research.

A. Camera-LiDAR fusion-based deep learning

The fusion of camera and LiDAR data stands out as one of the extensively investigated topics in multimodal fusion, particularly in the context of traffic object detection and segmentation. Various taxonomies are employed to categorize deep fusion algorithms that integrate camera and LiDAR information. To distinguish different fusion principles we adopt the patterns suggested in [9], namely *signal-level*, *feature-level*, *result-level*, and *multi-level* fusion. This systematic categorization aids in better understanding and comparing the diverse approaches employed in the fusion of camera and LiDAR data for enhanced performance in traffic-related applications.

(i) The ***signal-level*** fusion is expressed as early-stage fusion as it relies on spatial coordinate matching and raw data (e.g. 2D/3D geometric coordinates, image pixel values) integration to achieve the fusion of two sensing modalities. Depth completion [20] [21] is an iconic application which is instinctively suitable for *signal-level* fusion. Work [22] [23], and [24] explored the possibility of using *signal-level* fusion in road/lane detection scenarios and its performance-computation trade-off. There are relatively few works that implement *signal-level* fusion for traffic object detection and segmentation [25] [26] because texture information loss is inevitable in sparse mapping and projection process.

(ii) On the other hand, the literature of ***feature-level*** fusion is rich. In general, the LiDAR data is involved in fusion as either a voxel grid or 2D projection, and the feature map is the most common format for image input. VoxelNet [27] is the leading work to sample raw point clouds as sparse voxels before the fusion with camera data. The examples of the fusion of LiDAR's 2D projections and camera images are [28] [29] [30].

(iii) The intuition of ***result-level*** fusion is using the weight-based logical operations to combine the prediction results from different modalities, which is adopted in work [31] [32].

(iv) The ***multi-level*** fusion combines the other three fusion approaches mentioned above to overcome the shortcomings of the respective method. Van Gansbeke et al. [33] combined *signal-level* and *feature-level* fusion in a network for depth prediction. PointFusion [34] explored the *result-level* and *feature-level* fusion combination by first generating 2D bounding boxes, then filtering the LiDAR points based on these 2D boxes, at last, using a ResNet [35] and PointNet [7] network to integrate image and point clouds features to 3D object predictions. Other *multi-level* fusion research includes [36] [37].

During the literature review, we observe that the transition from *signal/result-level* to *multi-level* fusion is the general trend of camera-LiDAR deep fusion. To mitigate some limitations such as computational complexity, early works usually extract geometric information directly from LiDAR data to leverage the existing ready-to-use image processing networks. The recent research tends to carry out the fusion in a *multi-level* format, that adopts various fusion strategies and context encoding processes. Our work contributes in the line of a *multi-level* fusion architecture which uses a transformer head to encode the input and then execute the cross-fusion of camera and LiDAR data.

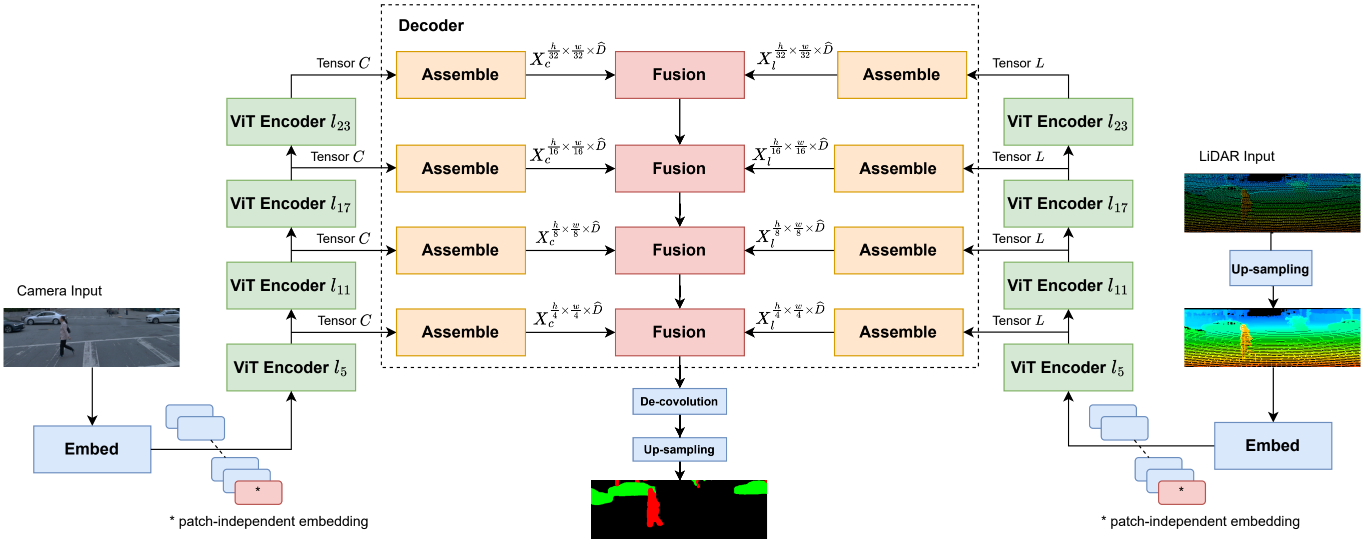
Decoder
Tensor C
Tensor L
ViT Encoder l_{23}
ViT Encoder l_{17}
ViT Encoder l_{11}
ViT Encoder l_{5}
Assemble
Fusion
$X_c^{\frac{h}{32}\times\frac{w}{32}\times\hat{D}}$
$X_l^{\frac{h}{32}\times\frac{w}{32}\times\hat{D}}$
$X_c^{\frac{h}{16}\times\frac{w}{16}\times\hat{D}}$
$X_l^{\frac{h}{16}\times\frac{w}{16}\times\hat{D}}$
$X_c^{\frac{h}{8}\times\frac{w}{8}\times\hat{D}}$
$X_l^{\frac{h}{8}\times\frac{w}{8}\times\hat{D}}$
$X_c^{\frac{h}{4}\times\frac{w}{4}\times\hat{D}}$
$X_l^{\frac{h}{4}\times\frac{w}{4}\times\hat{D}}$
Camera Input
LiDAR Input
Up-sampling
Embed
De-covolution
Up-sampling
*
* patch-independent embedding

Fig. 1. The overall architecture of our double-direction network shows camera data flowing from the left side into the ViT encoder, while LiDAR data flows from the right. The camera input is individual RGB channels, and the LiDAR input stands as XY, YZ, and XZ projection planes. The cross-fusion strategy is shown in the center and highlighted using a dashed rectangle.

B. Transformers in autonomous driving research

The attention mechanism [11] has garnered significant attention from researchers across diverse fields since its introduction by Vaswani et al. in the transformer architecture for natural language processing (NLP) tasks [10]. Among the most notable transformer variants is the Vision Transformer (ViT) [12], showcasing its capabilities in computer vision with direct applications in autonomous driving. Specifically, the autonomous driving perception tasks benefit the most from the attention-mechanism's strengths in global context and long-range dependencies handling. In this section, we review the state-of-the-art transformer-based works for 2D and 3D general perception in autonomous driving.

The 2D perception applications of autonomous driving extract the information from camera images. Lane detection is the most prevalent task among 2D perception research. Peng et al. [38] proposed a bird's eye view transformer-based architecture for road surface segmentation. Work [39] adopted a lightweight transformer structure for lane shape prediction, first modeled lane markings as regressive polynomials, then optimized the polynomial parameters by a transformer query and Hungarian fitting loss algorithms. Other transformer deep networks for road/lane segmentation include [15] [40]. There are relatively fewer works of 2D segmentation because the multimodal fusion is the trend for semantic segmentation in recent. Panoptic SegFormer [41] proposed a panoptic segmentation framework utilizing a supervised mask decoder and a query decoupling method to execute the semantic and instance segmentation.

The research of transformer-based 3D object detection and segmentation is abundant. DETR3D [42] is a variant of the popular DETR [43] model but extended its 2D object detection potential to 3D detection scenarios. DETR3D relied on multi-view images to recover 3D information and used backward geometric projection to combine 2D feature extraction and 3D prediction. FUTR3D [44] is a counterpart network to DETR3D, featuring a modality-agnostic feature sampler designed to accommodate multimodal sensory input for precise 3D bounding box predictions. PETR [45] embedded 3D coordinate information into image to produce 3D position-aware features. BEVFormer [46] employed spatial and temporal attention layers for bird's eye view features to improve the performance of 3D object detection and map segmentation. Work [47] and [48] focused on the 3D segmentation. TPVFormer [47] reduced the computational requirement by transforming the volume to three bird's eye view planes. VoxFormer [48] generated 3D voxels from 2D images, then performed cross and self attention mechanisms to 3D voxel queries to compute semantic segmentation results.

With reference to our review, there are relatively few research works on the semantic object segmentation, let alone the multimodal fusion of camera and LiDAR sensors. Work [44] and [16] directly used LiDAR input, but their focus are 3D detection and occupancy prediction. Moreover, other latest works [47] and [48] produced the voxel and pseudo-point-clouds from the camera input, then carried out the semantic occupancy prediction. While our CLFT models directly take LiDAR data as input, and adopt another strategy to process the LiDAR point clouds as image views in camera plane to achieve 2D semantic object segmentation. Foremost, our work plays a crucial role in bridging the gap in multimodal semantic object segmentation within the realm of autonomous driving research.

III. Methodology

There are two aims of our CLFT models in this work; first is to outperform the existing state-of-the-art single modality transformer-based models; second is to compete with the recent CNN-based models in terms of traffic object segmentation by fusing the camera and LiDAR data. We maintain the overall structure of the transformer network for dense prediction (DPT) [49] but invoke a late fusion strategy in its convolutional decoder, which first assemble the LiDAR and camera data in parallel and then integrate their feature map representations. We explore the capability of transformer-based networks in semantic segmentation with the advantages of LiDAR sensors, prove transformer networks' potential to classify the less represented samples in contrast with CNNs, at last, provide a late fusion strategy for transformer-related sensor fusion research.

The encoder-decoder structure has been widely implemented in image analysis transformers. We closely follow the protocol of ViT [12] to establish the encoders in our network to create the multi-layer perceptron (MLP) heads for camera and LiDAR data separately. For the decoders, we refer, but leverage proposals in work [49] to assemble and integrate the feature representations from camera and LiDAR sensors to create the object segmentation that is more precise than single modality. Figure 1 shows the overall architecture of our network.

1) Encoder: ViT innovatively proposed an encoder to convert an image into multiple tokens that can be treated in the same way as words in a sentence; consequently, transferred the standard transformer from NLP to computer vision applications. The ViT encoder uses two different procedures to transfer the images into tokens. The first approach divides an image into fixed-size non-overlapping patches, followed by linear projection of their flattened vector representations. The second approach extracts feature patches from a CNN feature map and then feeds them into the transformer as tokens. We retain the ViT's conventions to define the encoder variants in our work, namely, 'CLFT-base', 'CLFT-large', 'CLFT-huge', and 'CLFT-hybrid'. The 'base', 'large', and 'huge' indicate the encoder's configuration such as layer, size, and amount of parameters. The 'hybrid' means other neural network backbones are integrated in the model. The 'CLFT-base', 'CLFT-large', and 'CLFT-huge' architectures use patch-based embedding methods, have 12, 24, and 32 transformer layers, and the feature dimension D of each token are 768, 1024, and 1280, respectively. The 'CLFT-hybrid' encoder employs a ResNet50 network to extract pixel features as image embeddings, followed by 12 transformer layers. The patch size p of all our experiments is 16. The resolution of the input camera and LiDAR image (h, w) is (384, 384), which means the total amount of pixels for each patch $\frac{h*w}{p^2} = 576$ is smaller than feature dimensions D of all variants; thus, the knowledge can be retrieved from input in pixel-wise. For the 'CLFT-hybrid' encoder, it extracts the features from the input patch of $384 \div 16 = 24$ resolution. All the encoders are pretrained using ImageNet [50]. Following work in ViT, we concatenate position embeddings with image embeddings to retain positional information. Moreover, there is an individual learnable token in sequence for classification purposes. This classification token is represented as red block with the asterisk in Figure 1. It is similar to BERT's 'class' token [51], independent from all image patches and positionally embedded. Please refer to the original work [12] for the details of these encoder architectures.

2) Decoder: The transformer networks designed for computer vision usually modify the decoder by implementing convolutional layers at different stages. Ranftl et al. [49] proposed a transformer network for dense prediction (DPT) that progressively assembles tokens from various encoder layers into image-like representations to achieve final dense prediction. Inspired by DPT's decoder architecture, we construct a decoder to process the LiDAR and camera tokens in parallel.

As illustrated in Figure 1, we pick four transformer encoder layers denoted as t ($t = \{2, 5, 8, 11\}$ for 'CLFT-base' and 'CLFT-hybrid', $t = \{5, 11, 17, 23\}$ for 'CLFT-large'), then assemble the tokens from each layer to an image-like representation of feature maps. The feature map representations at the initial layers of the network are up-sampled to a high resolution, whereas representations from deep layers ware down-sampled to a low resolution. The resolutions are anchored to input image size (h, w), and the sampling coefficients corresponding to encoder layers t are $s = \{4, 8, 16, 32\}$. In detail, there are two steps in the assembly process. As illustrated in Algorithm 1, the first step replicates and concatenates the patch-independent 'classification token' with all other tokens individually, then forwards the concatenated representations to an MLP process with GELU non-linear activation [52]. The number of individual tokens is denoted as k.

Algorithm 1 The projection of the 'classification token'.

```
Input: Input tensor $T$, representing either the camera or
    LiDAR channels containing the 'classification token' and
    patch tokens.
Output: Concatenated tensor representations $X_T$
1: $T_{cls} = replicate\{T[:, 0]\}$
2: $T_{concat} = T[:, i] \mathbin\Vert T_{cls} \quad \forall \ i = 1, \ldots, k$
3: $X_T = \text{GELU}(W \cdot T_{concat} + b)$
```

Equation 1 shows the second step, which first concatenates the tokens from the first step based on their initial positional order to yield an image-like representation, then passes this representation to two convolution operations. The first convolution projects the representation from dimension D to $\hat{D}$ ($\hat{D}$ is set as 256 in our experiments). The second convolution applies up-sampling and down-sampling toward representation concerning the different layers of transformer encoders. X_c and X_l are the concatenated camera and LiDAR representations, N represents the total amount of patches. The generic workflows of these two steps are shown in Figure 2.

$$X_t^{N \times D} \Rightarrow X_t^{\frac{h}{s} \times \frac{w}{s} \times \hat{D}} \tag{1}$$

$$X_t = \{X_c, X_l\} \quad s = \{4, 8, 16, 32\}$$

$$t = \{2, 5, 8, 11\} \ or \ \{5, 11, 17, 23\}$$

The last process of our decoder is the cross-fusion of camera and LiDAR feature maps, which is progressively illustrated in Figure 3. We refer to the feature fusion strategy from RefineNet [53] that forwards the camera and LiDAR representations through two residual convolution units (RCU) in sequence. The camera and LiDAR's representations are

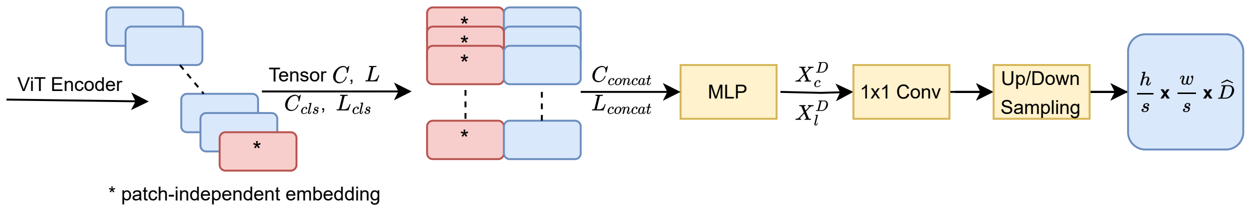

Fig. 2. Assemble architecture for each transformer decoder block, tokens of each layers are assembled to image-like representations of feature maps.

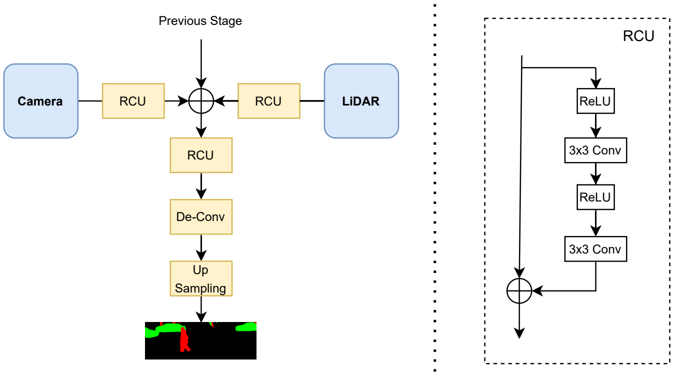

Fig. 3. Each fusion block receives data from the previous stage and integrates camera-LiDAR data coming from the ViT encoder. Each of this block has residual units, de-convolution, and up-sampling.

summed with the results from the previous fusion operation and then went through one additional RCU. We pass the output of the last fusion layer to a deconvolutional and up-sampling module to compute the final predicted segmentation. The fusion of the information coming from the LiDAR and the camera can happen in any of the fusion block as the connection weights are automatically learned in the network through error back-propagation. The idea of our multiple fusion blocks is to integrate the concept of late-fusion (as each fusion blocks is placed after each assemble block) and the concept of cross-fusion [24] as the connection with each feature map can happen in any of the fusion blocks with different weights. The network automatically learns to weight the best block to integrate tensor information coming from different sensors.

IV. Dataset Configuration

The primary purpose of this work is to compare the performance of the vision transformer and CNN backbones for semantic segmentation. Our previous work [13] successfully modeled and evaluated a ResNet50-based FCN to carry out camera-LiDAR fusion. In order to maintain an accordant experiment environment, we construct the input data based on Waymo dataset [54] to evaluate CLFT and other models.

Waymo dataset is recorded by multiple high-quality cameras and LiDAR sensors. The scenes of Waymo dataset span various illumination levels, weather conditions, and traffic scenarios. Therefore, as shown in Table I, we manually partitioned the data sequences into four subsets: light-dry, light-wet, dark-dry, and dark-wet. The ‘light’ and ‘dark’ indicate the relative illumination conditions. The ‘dry’ and ‘wet’ represent the weather difference in precipitation.

TABLE I
Amount of the Frames in Four Broad Subsets for Waymo Open Dataset.

Light-Dry	Dark-Dry	Light-Wet	Dark-Wet
14940	1640	4520	900

We provide intersection over union (IoU) as the primary indication of model evaluation, with precision and recall values as supplementary information. Please note that the IoU is primarily used in object detection applications, in which the output is the bounding box around the object. Therefore, We modify the ordinary IoU algorithm to fit the multi-class pixel-wise semantic object segmentation. The essential change is related to the ambiguous pixels (pixels have no valid labels, details in Section IV-B) that fall out of the class list. We assign these pixels as void and exclude them from the evaluation. The performance of networks is measured by the statistics of the number of pixels that have identical classes indicated in prediction and ground truth.

A. LiDAR Data Processing

The LiDAR readings reflect the object’s 3D geometric information in the real world. Coordinate values in three spatial channels contain features that can be exploited by neural networks. As a result, regarding camera-LiDAR fusion, it is common to extract and fuse multi-target features such as images’ color textures and point clouds’ location information, which is an approach namely as feature-level fusion [55].

We adopt feature-level fusion in this work. Thus, we project 3D LiDAR point clouds into the camera plane to create 2D occupancy grids in XY, YZ, and XZ planes. All the points in LiDAR point clouds are transformed and projected following Equation 2 and 3, respectively.

$$\left[x_t, y_t, z_t\right]^T = \left(r\ p\ y\right)\left(\left[x_i, y_i, z_i\right]^T - \left[x_c, y_c, z_c\right]^T\right) \quad (2)$$

$$r = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\rho) & sin(\rho) \\ 0 & -sin(\rho) & cos(\rho) \end{bmatrix} p = \begin{bmatrix} cos(\theta) & 0 & -sin(\theta) \\ 0 & 1 & 0 \\ sin(\theta) & 0 & cos(\theta) \end{bmatrix} y = \begin{bmatrix} cos(\phi) & sin(\phi) & 0 \\ -sin(\phi) & cos(\phi) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

In Equation 2, x_t, y_t, and z_t are the 3D point coordinates after transformation (in camera frame); r, p, and y represent the Euler rotation matrices to the camera frame with (ρ, θ, ϕ) representing the corresponding Euler angles. x_i, y_i, and z_i are the 3D point coordinates before transformation (in LiDAR

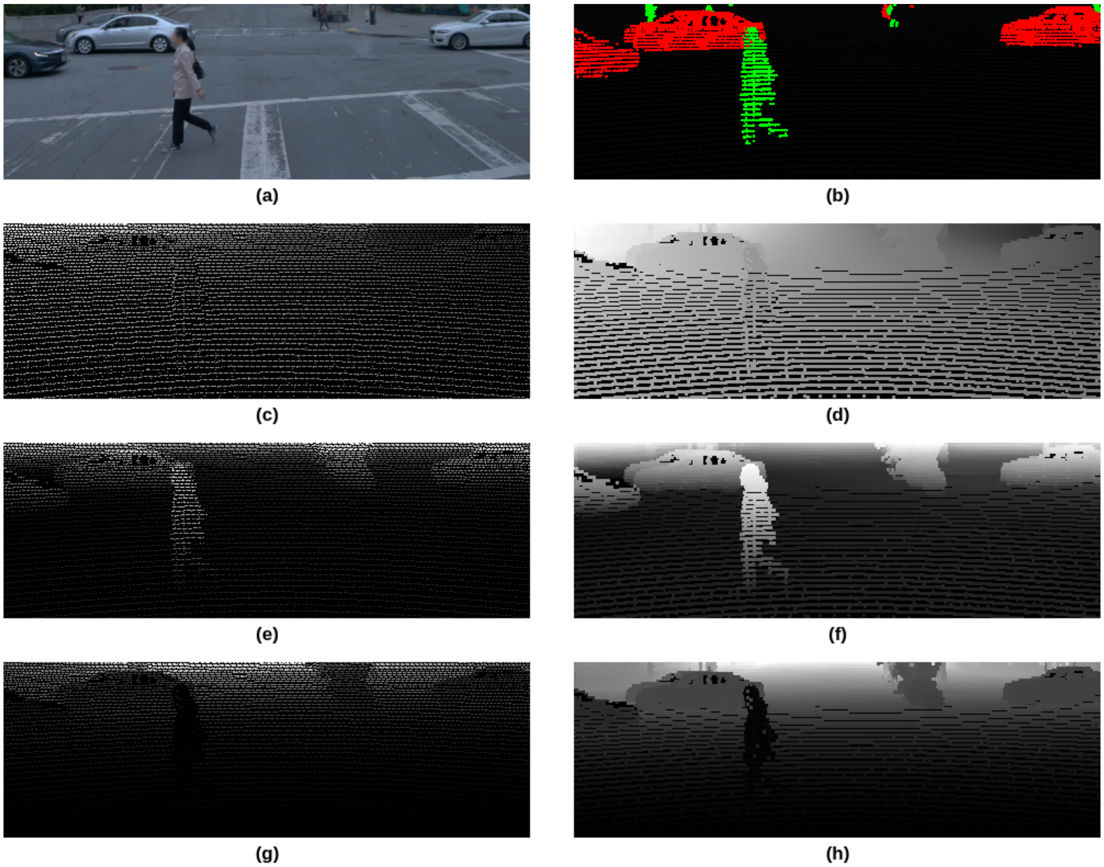

Fig. 4. Examples of camera image, semantic annotation mask, and pre-processing of LiDAR data. (a) is the RGB image. (b) illustrates the object semantic masks obtained from LiDAR ground truth bounding boxes. (c) (e) (g) are LiDAR projection images in X, Y, Z channels, respectively, while (d) (f) (h) are corresponding up-sampled dense images. Please note that for visualization purposes, the grayscale intensity in (c)-(h) is proportionally scaled based on the numerical 3D coordinate values of the LiDAR points.

frame); x_c, y_c, and z_c denote the camera frame location coordinates.

$$(u, v, 1)^T = \begin{bmatrix} f_x & 0 & \frac{w}{2} \\ 0 & f_y & \frac{h}{2} \\ 0 & 0 & 1 \end{bmatrix} (x, y, z)^T \tag{3}$$

In Equation 3, u and v are column and row positions of the point in 2D image plane; f_x and f_y denote camera's horizontal and vertical focal length; w and h represent image resolution; x, y, and z are transformed 3D point coordinates (same as x_t, y_t, and z_t in Equation 2).

Algorithm 2 LiDAR points filtering and image pixel values population

```
Input: LiDAR point 3D coordinates $L$, projected LiDAR point coordinates $P$, image resolution $w$ and $h$.
Output: LiDAR projection footprints $XY$, $YZ$, and $ZX$.
$idx = argwhere(P < \{w, h, +\infty\} \ \& \ P >= \{0, 0, 0\})$
$XY[w \times h] \leftarrow 0$
$YZ[w \times h] \leftarrow 0$
$XZ[w \times h] \leftarrow 0$
$XY[idx] = L[idx, 0]$
$YZ[idx] = L[idx, 1]$
$XZ[idx] = L[idx, 2]$
```

The operation after transforming and projecting the 3D point clouds into 2D images is filtering, which aims to discard all the points that fall out of the camera view. Waymo Open dataset is collected using five LiDAR and five camera sensors covering all vehicle directions. This work uses the top LiDAR's point clouds and the front camera's image data. As shown in Algorithm 2, three projection footprint images denoted as XY, YZ, and ZX are generated. The pixels corresponding to 3D points are assigned with x, y, and z coordinates, while the rest are populated with zero. At last, we up-sample the LiDAR images before feeding them to machine learning algorithms, as it is a common practice in LiDAR-based object detection research [56] [57]. Figure 4 (c)-(g) show the results of the procedure described in this subsection.

B. Object Semantic Masks

Ground truth annotations in Waymo dataset are represented by 2D and 3D bounding boxes, which correspond to camera and LiDAR data separately. There are three classes in image annotations: vehicles, pedestrians, and cyclists. Point clouds annotations have an extra class which is traffic signs. There are two obstacles when using Waymo's ground truth annotations in our networks.

Firstly, vision-transformer-based networks are well-known for requiring vast samples [12]. However, the cyclists and

TABLE II
PERFORMANCE COMPARISON OF CLFT-HYBRID VARIANT, CLFCN AND PANOPTIC SEGFORMER. BOLD INDICATES THE BEST VALUES IN EACH ROW PER CLASS. (IN PERCENTAGE UNIT) (C, L, AND C+L INDICATE CAMERA-ONLY, LIDAR-ONLY, AND FUSION MODALITIES, RESPECTIVELY)

	CLFT-Hybrid		CLFCN		CLFCN		CLFCN		Panoptic SegFormer		Panoptic SegFormer	
	(C+L)		(C)		(L)		(C+L)		(C)		(L)	
	Vehicle	Human	vehicle	Human	Vehicle	Human	Vehicle	Human	Vehicle	Human	Vehicle	Human
Light-Dry	**91.35**	**66.04**	88.08	55.57	88.58	53.04	91.07	62.50	85.89	61.02	66.41	40.78
Light-Wet	91.72	**66.03**	88.54	52.13	89.47	50.06	**92.77**	64.66	83.58	49.70	63.07	29.87
Dark-Dry	**90.62**	**65.66**	81.16	42.87	86.16	48.83	89.41	60.33	81.45	44.67	70.25	38.69
Dark-Wet	**90.18**	53.51	74.49	43.14	87.51	46.68	89.90	**56.70**	70.50	14.68	54.40	39.00

traffic signs are relatively rare-represented in the Waymo dataset. We notice our CLFT models struggle to learn and predict these two classes in experimental setting as they are less represented in the dataset. We assume that with additional data also traffic signs and cyclists can be properly classified. Therefore, we discard the traffic signs in this work and merge the cyclists and pedestrians as a new class of so-called human.

Secondly, our research aims for semantic segmentation, which requires annotations denoted as object contours. Since Waymo dataset labeled the object in LiDAR sensor readings as a 3D upright bounding box, we project all the points in the bounding box into the image plane by the same procedure described in Section IV-A. Figure 4 (b) shows an example of semantic masks for vehicle and human classes. Please note that a limitation of this approach is that some object pixels have no valid labels because there are no corresponding LiDAR points.

V. RESULTS

As mentioned in Section I, our CLFT is the first transformer-based model fusing the camera and LiDAR sensory data for semantic segmentation. The experiments in this work focus on the controlled benchmark comparisons in two aspects: i) neural network architecture, ii) input modality.

The FCN is believed to be the recent generation of deep learning methods with remarkable performance improvements and has become the mainstream for semantic segmentation [58]. Therefore, we choose the CLFCN [13], an FCN-based network that fuses camera and LiDAR data for semantic segmentation, as the reference to explore the advantages of transformer backbone. Since the transformer is well-known for its strengths in capturing global context and solving long-range dependencies, we expect our transformer-based model to outperform the FCN-based model in scenarios such as unevenly distributed datasets and underrepresented samples.

Only a few existing deep learning methods process the LiDAR input using the same principle as in this work: representing the 3D point clouds as 2D grid-based feature maps [14]. We compare the CLFT with the Panoptic SegFormer [41] that is also transformer-backbone to evaluate the significance of various input modalities. However, the Panoptic SegFormer is purely vision-based. We follow the procedures in Section IV to produce the point clouds projection images as LiDAR modality input for Panoptic SegFormer, but the camera-LiDAR-fusion mode is not directly applicable to Panoptic SegFormer. It is critical to maintain the same input data splits and configurations in experiments for all models.

A. Experimental setup

The details of the input dataset configuration are described in Section IV. The dataset splits for training, validation, and testing are 60%, 20%, and 20% of the total number of frames, respectively. The four data subsets, light-dry, light-wet, dark-dry, and dark-wet, are shuffled and mixed for training and validation but tested individually. We adopt the default hyper-parameter configurations for CLFCN and Panoptic SegFormer in training. Please refer to authors' original work for details [41]. We employ weighted cross-entropy loss function and Adam optimization [59] for CLFT networks training. The transformer encoder of CLFT is initiated from ImageNet pre-trained weights. The transformer decoder and CLFCN's ResNet backbone are initiated randomly. The learning rate decay of CLFT networks training follows $l_i = l_0(\alpha^i)$, where l_0 is the initial learning rate, and α is 0.99. The batch size of CLFT networks training is set as 32 by default, but set as 24 for several experiments that exceed the memory limit, for example, the fusion mode of CLFT-large variant. Other hyperparameter settings can be found in the code we public. The transformer-based networks are trained using an NVIDIA A100 80GB GPU due to the large memory requirement of transformer networks. Relatively low-memory-required FCN training is executed on a desktop equipped NVIDIA RTX2070 Super GPU. The software environment of all experiments is Python3.9 and CUDA11.2. Please refer to our GitHub link for more details about the environment. Data normalization, augmentation and early stopping are also used to generate the models as in all most recent state-of-the-art methods.

B. Network performance and comparison

The main result of this work is reported in Table II and Table III. Values are shown as the IoU for the two interest classes, vehicle and human, in different modalities and weather scenarios. The modalities are indicated as C, L, and C+L, referring to the camera, LiDAR, and fusion, respectively.

As shown in Table II, the CLFT-hybrid variant outperforms the CLFCN and Panoptic SegFormer in all scenarios, demonstrating high segmentation capabilities over the same data. Specifically, in dry environmental conditions, CLFT-hybrid fusion modality archives 91% IoU for vehicles and 66% for humans, while CLFCN fusion modality has 90% for vehicles and 61% for humans. For single modality, Panoptic SegFormer achieves a similar performance of CLFCN for vehicle class but outperforms for human class with less fine-tuned works (61.02% against 55.57% in light-dry environment),

TABLE III
PERFORMANCE COMPARISON OF ALL CLFT VARIANTS, CLFCN, AND PANOPTIC SEGFORMER. (IN PERCENTAGE UNIT)(C, L, AND C+L INDICATE CAMERA-ONLY, LIDAR-ONLY, AND FUSION MODALITIES, RESPECTIVELY)

	VEHICLE			HUMAN		
	Precision	Recall	IoU	Precision	Recall	IoU
CLFT-Base (C+L)	93.63	95.95	90.12	71.97	79.47	60.68
CLFT-Large (C+L)	93.81	96.14	90.46	72.27	77.76	60.56
CLFT-Hybrid (C+L)	94.15	96.69	**91.26**	75.76	82.75	**65.46**
CLFCN (C+L)	93.17	**97.67**	91.19	65.63	**92.89**	62.51
Panoptic SegFormer (C)	**94.82**	88.43	84.40	**81.11**	63.78	55.55
Panoptic SegFormer (L)	89.57	70.85	65.48	67.84	46.85	38.29

which reinforces the transformer's strength regarding underrepresented samples. The difference between our CLFT and other models is even more evident in challenging conditions such as dark and wet, where CLFT-hybrid performance drops by 1-2 percentage points while CLFCN and Panoptic SegFormer in single modalities drop by 5-10 percentage points. In these cases, fusion seems to play a pivotal role in CLFCN while showing only slight improvements in CLFT-hybrid, demonstrating the robustness of CLFT-hybrid in performing data fusion in all types of conditions.

The Panoptic SegFormer has obvious weak performance in LiDAR modality. This is because it is designed to process RGB visual input. We carry out the LiDAR processing separately to produce the camera-plane maps with 3D coordinate information; then we feed the maps to Panoptic SegFormer. The experiment results prove the necessity to integrate the LiDAR processing into the neural networks' architecture. Though CLFT-hybrid outperforms the CLFCN in fusion in most cases, it is essential to see that CLFCN models benefit more from the fusion, as the improvement from individual modalities seems to be higher, particularly in night conditions. On the other hand, our CLFT models already show high performance in challenging conditions with the fusion of camera and LiDAR data.

Table III summarizes the performance of CLFT variants, CLFCN, and Panoptic SegFormer. We present the precision, recall, and IoU for all models. In order to have a straightforward comparison, we combine four weather scenarios for performance evaluation. In all cases, the CLFT-hybrid variant performs better than the base and huge variants. This result is consistent with what Dosovitskiy et al. [12] reported in their ablation experiments, in which ResNet-based transformer variants outperform the variants that use patch-based embedding procedures. Though the CLFT-hybrid achieves the highest IoU score, CLFCN and Panoptic SegFormer have higher recall and precision results, respectively.

C. Ablation study

Table IV reports our results using camera (C), LiDAR (L), and fusion (C+L). According to our ablation study in Table IV, it is possible to conclude that fusion provides an improvement over single-modality networks.

One might note that results for the individual modalities, particularly LiDAR, show already performance over 90% (before fusion); this result is also in line with many other studies in the field, for instance, in [60] the authors reached over 90% IoU in the car class on the SemanticKitti dataset [61].

TABLE IV
ABLATION STUDY BASED ON CLFT-HYBRID VARIANT. (IN PERCENTAGE UNIT)

(C, L, and C+L indicate camera-only, LiDAR-only, and fusion modalities, respectively)

C	L	IoU		Precision		Recall	
		Vehicle	Human	Vehicle	Human	Vehicle	Human
		All weather					
✓		91.16	64.38	93.86	73.33	96.88	84.05
	✓	91.19	65.17	93.93	72.89	96.85	84.19
✓	✓	**91.26**	**65.46**	94.15	75.76	96.69	82.75
		Light-Dry					
✓		91.23	64.87	93.83	72.63	97.05	85.86
	✓	91.32	64.92	93.96	72.68	97.02	85.88
✓	✓	**91.35**	**66.04**	94.14	75.31	96.86	84.29
		Light-Wet					
✓		91.67	64.87	94.52	76.49	96.82	81.36
	✓	91.52	64.28	94.40	74.43	96.78	82.49
✓	✓	**91.72**	**66.03**	94.69	78.27	96.96	80.84
		Dark-Dry					
✓		90.51	65.62	93.15	74.30	96.96	84.66
	✓	90.47	65.18	93.27	74.30	96.96	84.16
✓	✓	**90.62**	**65.66**	93.38	77.39	96.68	81.25
		Dark-Wet					
✓		89.62	52.46	93.60	70.00	95.70	67.69
	✓	89.74	49.95	93.69	67.28	95.51	65.97
✓	✓	**90.18**	**53.51**	94.40	68.68	95.29	70.79

Inspecting the analysis on all-weather, one can see that CLFT-hybrid provides a small improvement (less than one percentage point in both classes). However, as by construction, the dataset split is strongly unbalanced (see Table I) toward light-dry scenario (roughly 68% of the total). The amount of light scenarios covers over 88% of the total number of frames. Clearly, the class that is better represented in the dataset affects the overall result the most.

To better appreciate the improvement in our studies, Table IV is also divided according to the data split in Table I. Under these conditions, it is possible to assert that fusion has a higher impact in dark scenarios, covering roughly 12% of the total number of frames in our dataset.

The unbalance of the dataset has an impact on both environment conditions and object classes, thus the vehicle class (with already over 90% accuracy) is less affected, while the human class shows better improvements, reaching around 2-4% in rainy conditions.

D. Inference time analysis

Table V presents an additional study on the inference time. In the experiments, we make the statistic of CUDA event time on NVIDIA A100 GPU for fusion modality of all models. All the models are set in evaluation mode for inference time calculation. We use the image in Figure 4 as input, first warm up the GPU with 2000 iterations, then calculate the mean time of the event stream for another 2000 iterations. The CPU and GPU are synchronized when recording timestamps. In general, FCN-based models have obvious advantages against the transformer-based models in terms of computational efficiency. The Panoptic SegFormer has the highest inference time among all models in experiments. It appears that the CLFCN is faster than our best-performing model, the CLFT-hybrid. However, this difference is only about 10ms per frame, which can be considered reasonable in a trade-off between performance and speed. For autonomous driving, where safety comes first, classification performance should always be considered a crucial parameter in the network design.

TABLE V
INFERENCE TIME COMPARISON OF ALL CLFT VARIANTS, CLFCN AND PANOPTIC SEGFORMER (IN MILLISECONDS UNIT)(C, L, AND C+L INDICATE CAMERA-ONLY, LIDAR-ONLY, AND FUSION MODALITIES, RESPECTIVELY)

NETWORK	MODALITY	TIME
CLFT-base	C+L	16.23
CLFT-Large		36.75
CLFT-Hybrid		25.69
CLFCN		15.94
Panoptic SegFormer	C	93.52
	L	93.45

E. Qualitative results

Figure 5 presents examples of segmented images from the Waymo dataset to appreciate the results of this work from a qualitative point of view. Following the above mentioned contribution of this work, the qualitative evaluation is also divided by network structure, weather and illumination conditions. The three CLFT variants, ‘Base’, ‘Large’, and ‘Hybrid’, are compared with the Panoptic SegFormer and CLFCN modalities. The segmentation results from models are overlaid to the camera images for comparison. The first row is the ground truth segmentation provided by the dataset. Please note that the annotations of the Waymo dataset are based on the LiDAR point clouds data, which is a common labeling strategy adopted by many famous multi-modal datasets for autonomous driving, including SemanticKitti and nuScenes [62] datasets. The LiDAR-points-based labeling strategy results the 2D semantic masks contain the pixels without valid label. Waymo dataset claimed to have the highest per-frame point clouds density among the SemanticKitti, nuScenes, and Argoverse [63] datasets, which is the reason why the Waymo dataset better fits for the evaluation of CLFT networks for 2D semantic segmentation tasks.

The qualitative results generally follow the same consistency as in numerical benchmarks. The CLFT-Hybrid variant discloses the most contextual details and its segmentation results are more identical to ground truth than other networks, especially in challenging and under-represented environments. For example, the vehicles in night-dry (the third column) scenario, the CLFCN networks detect less details even with fine-tuning efforts, proves that the transformer is more effective than FCN in specific situations. Moreover, the single-modality segmentation results from Panoptic SegFormer and CLFCN networks show the necessities and advancements of multi-modal sensor fusion in autonomous driving.

VI. CONCLUSION

In this paper, we propose a transformer-based multimodal fusion method for semantic segmentation. Based on all the above cases, it is possible to say our CLFT model is one of the cutting-edge neural networks for 2D traffic object semantic segmentation. Specifically, the CLFT models benefit from the multimodal sensor fusion and transformer’s multi-attention mechanism, make a significant improvement for under-represented samples (maximum 10 percent IoU increase for human class). However, it is worth mentioning that transformer networks intuitively require a large amount of data for training. In our experiments, light-wet and dark-wet subsets only take into account 12% of the total input data, which explains that the CLFCN model outperforms the CLFT-hybrid model in some cases in Table II.

This work proposes the adoption of a vision transformer’s strategy to divide the input image into non-overlapping patches or extract feature patches from CNN feature maps. Intuitively, we project and up-sample LiDAR data to dense point clouds images, then design a double-direction network to assemble and cross-fuse the camera and LiDAR representations to achieve final segmentation. We maintain the same input dataset splits and configurations in all our experiments and successfully demonstrate the transformer’s merit against the FCN regarding object segmentation tasks. Specifically, we classify the input data into sub-categories of different illumination and weather conditions dedicated to comprehensively evaluating the models. Similar to prior transformer works, we prove its potential on uneven-distributed datasets and under-represented samples. At last, we want to highlight that the initiation of CLFT lies on the progress to extend our framework that aims to cover all aspects of low-speed autonomous shuttles, including hardware configuration, dataset collection and post-processing for perception [17], validation [64], and path planning [65]. We develop the CLFT to be compatible with other systems in terms of environment, data formats, and operating platforms, which grants our work the advantages in scalability and practical application on real autonomous shuttles.

ACKNOWLEDGMENT

This research was funded by the European Union’s Horizon 2020 Research and Innovation Programme, under the grant agreement No. 856602. This research was co-funded by the European Union under the project Robotics and advanced industrial production (reg. no. CZ.02.01.01/00/22_008/0004590).

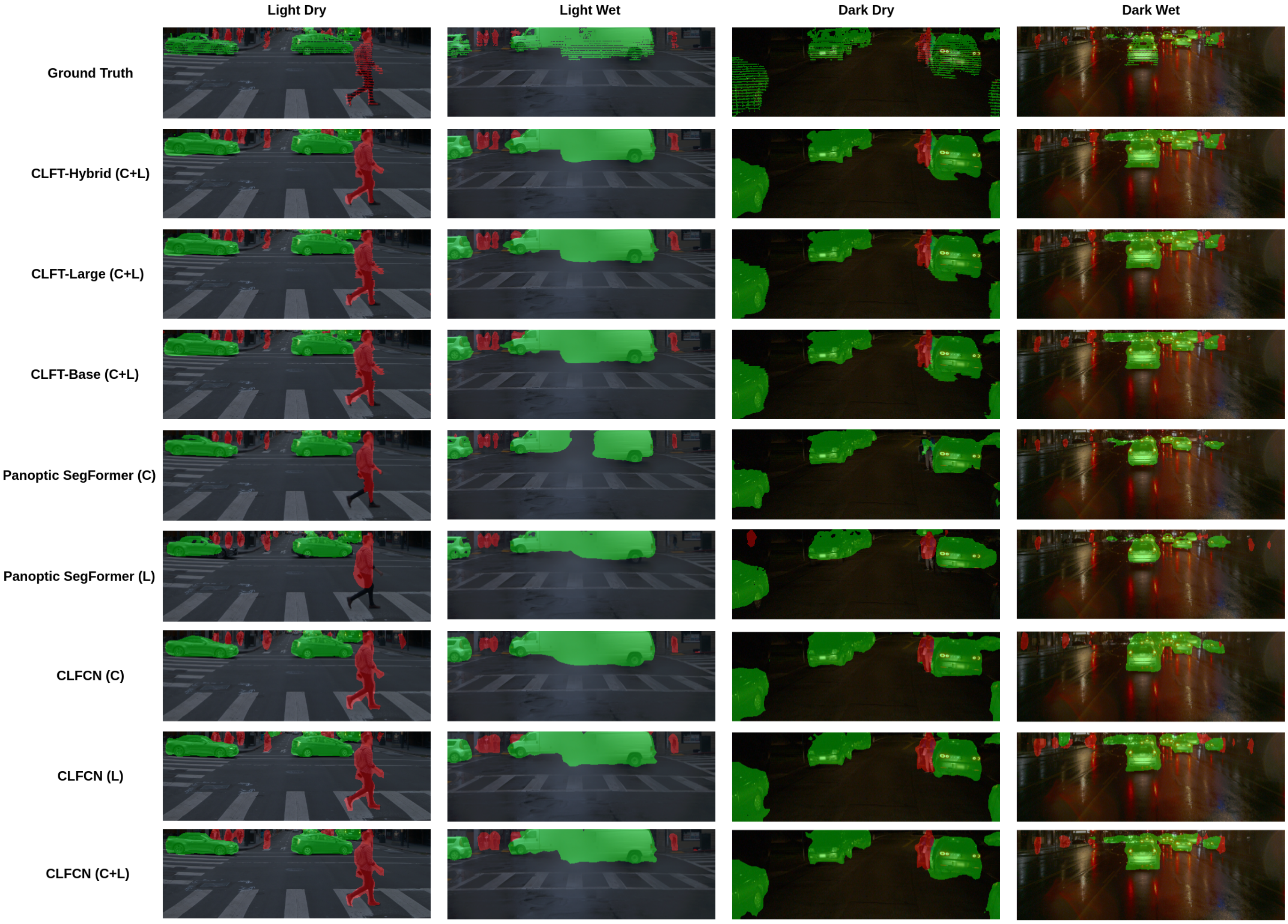

Fig. 5. Qualitative comparison of segmentation results between different models.

References

[1] L. Bartolomei, L. Teixeira, and M. Chli, "Perception-aware path planning for uavs using semantic segmentation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5808–5815.

[2] D. K. Dewangan and S. P. Sahu, "Driving behavior analysis of intelligent vehicle system for lane detection using vision-sensor," *IEEE Sensors Journal*, vol. 21, no. 5, pp. 6367–6375, 2021.

[3] J. Fritsch, T. Kühnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, 2013, pp. 1693–1700.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[6] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 922–928.

[7] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[8] A. Kanezaki, Y. Matsushita, and Y. Nishida, "Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5010–5019.

[9] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li, and D. Cao, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722–739, 2021.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] J. Gu, M. Bellone, R. Sell, and A. Lind, "Object segmentation for autonomous driving using iseauto data," *Electronics*, vol. 11, no. 7, p. 1119, 2022.

[14] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[15] L. Chen, C. Sima, Y. Li, Z. Zheng, J. Xu, X. Geng, H. Li, C. He, J. Shi, Y. Qiao *et al.*, "Persformer: 3d lane detection via perspective transformer and the openlane benchmark," in *European Conference on Computer Vision*. Springer, 2022, pp. 550–567.

[16] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 442–18 455, 2022.

[17] J. Gu, A. Lind, T. R. Chhetri, M. Bellone, and R. Sell, "End-to-

end multimodal sensor dataset collection framework for autonomous vehicles," *Sensors*, vol. 23, no. 15, 2023.

[18] J. Zhong, Z. Liu, and X. Chen, "Transformer-based models and hardware acceleration analysis in autonomous driving: A survey," 2023.

[19] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic segmentation using vision transformers: A survey," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106669, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197623008539

[20] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 3288–3295.

[21] X. Cheng, P. Wang, C. Guan, and R. Yang, "Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 615–10 622.

[22] F. Wulff, B. Schäufele, O. Sawade, D. Becker, B. Henke, and I. Radusch, "Early fusion of camera and lidar for robust road detection based on u-net fcn," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1426–1431.

[23] J.-S. Lee and T.-H. Park, "Fast road detection by cnn-based camera–lidar fusion and spherical coordinate transformation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 9, pp. 5802–5810, 2021.

[24] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "Lidar–camera fusion for road detection using fully convolutional neural networks," *Robotics and Autonomous Systems*, vol. 111, pp. 125–131, 2019.

[25] J. Dou, J. Xue, and J. Fang, "Seg-voxelnet for 3d vehicle detection from rgb and lidar data," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4362–4368.

[26] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

[27] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[28] L. Caltagirone, M. Bellone, L. Svensson, M. Wahde, and R. Sell, "Lidar–camera semi-supervised learning for semantic segmentation," *Sensors*, vol. 21, no. 14, p. 4813, 2021.

[29] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: A simple and robust lidar-camera fusion framework," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 421–10 434, 2022.

[30] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1239–1285, 2020.

[31] M. Jaritz, T.-H. Vu, R. d. Charette, E. Wirbel, and P. Pérez, "xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 605–12 614.

[32] S. Gu, T. Lu, Y. Zhang, J. M. Alvarez, J. Yang, and H. Kong, "3-d lidar+ monocular camera: An inverse-depth-induced fusion framework for urban road detection," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 3, pp. 351–360, 2018.

[33] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool, "Sparse and noisy lidar completion with rgb guidance and uncertainty," in *2019 16th International Conference on Machine Vision Applications (MVA)*, 2019, pp. 1–6.

[34] D. Xu, D. Anguelov, and A. Jain, "Pointfusion: Deep sensor fusion for 3d bounding box estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 244–253.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] K. El Madawi, H. Rashed, A. El Sallab, O. Nasr, H. Kamel, and S. Yogamani, "Rgb and lidar fusion based 3d semantic segmentation for autonomous driving," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 7–12.

[37] X. Zhao, Z. Liu, R. Hu, and K. Huang, "3d object detection using scale invariant and feature reweighting networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9267–9274.

[38] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5935–5943.

[39] R. Liu, Z. Yuan, T. Liu, and Z. Xiong, "End-to-end lane shape prediction with transformers," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3694–3702.

[40] Y. Bai, Z. Chen, Z. Fu, L. Peng, P. Liang, and E. Cheng, "Curveformer: 3d lane detection by curve propagation with curve queries and attention," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 7062–7068.

[41] Z. Li, W. Wang, E. Xie, Z. Yu, A. Anandkumar, J. M. Alvarez, P. Luo, and T. Lu, "Panoptic segformer: Delving deeper into panoptic segmentation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1280–1289.

[42] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*. PMLR, 2022, pp. 180–191.

[43] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[44] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 172–181.

[45] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 531–548.

[46] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022, pp. 1–18.

[47] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9223–9232.

[48] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. Anandkumar, "Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9087–9098.

[49] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[51] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[52] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[53] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[54] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[55] K. Banerjee, D. Notz, J. Windelen, S. Gavarraju, and M. He, "Online camera lidar fusion and object detection on hybrid data for autonomous driving," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 1632–1638.

[56] C. Premebida, J. Carreira, J. Batista, and U. Nunes, "Pedestrian detection combining rgb and dense lidar data," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 4112–4117.

[57] J. Schlosser, C. K. Chow, and Z. Kira, "Fusing lidar and images for pedestrian detection using convolutional neural networks," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 2198–2205.

[58] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," *Neurocomputing*, vol. 493, pp. 626–646, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222000054

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106

[60] X. Zhu, H. Zhou, T. Wang, F. Hong, W. Li, Y. Ma, H. Li, R. Yang, and D. Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar-based perception," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6807–6822, 2021.

[61] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences," in *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.

[62] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[63] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8748–8757.

[64] M. Malayjerdi, Q. A. Goss, M. İ. Akbaş, R. Sell, and M. Bellone, "A two-layered approach for the validation of an operational autonomous shuttle," *IEEE Access*, 2023.

[65] E. Malayjerdi, R. Sell, M. Malayjerdi, A. Udal, and M. Bellone, "Practical path planning techniques in overtaking for autonomous shuttles," *Journal of Field Robotics*, vol. 39, no. 4, pp. 410–425, 2022.

Gu Junyi received the B.S. degree in School of Optical-Electrical and Computer Engineering from the University of Shanghai for Science and Technology, Shanghai, China, in 2017. He received the M.S. degree in the Institute of Technology from the University of Tartu, Tartu, Estonia, in 2020. He is currently pursuing the Ph.D. degree at the Department of Mechanical and Industrial Engineering, Tallinn University of Technology, Tallinn, Estonia. His research interests include multi-sensor fusion, semantic segmentation, artificial intelligence, and autonomous driving.

Mauro Bellone received his M.S. degree in Automation Engineering from the University of Salento, Lecce, Italy, where he received his Ph.D. in Mechanical and Industrial Engineering in 2014. His interests comprise mobile robotics, autonomous vehicles, energy, computer vision, and control systems. His research focuses on advanced sensory perception for mobile robotics and artificial intelligence. From 2015 to 2020, he worked with the applied artificial intelligence research group of Chalmers University of Technology, where he actively contributed to several autonomous driving projects. In 2021, he was appointed as an adjunct professor at Tallinn University of technology, supporting the research team in the area of smart transportation systems.

Tomáš Pivoňka has received his master's degree in robotics at Faculty of Electrical Engineering of Czech Technical University in Prague (CTU) in 2018, where he continues in Ph.D. study program Artificial Intelligence and Biocybernetics. He works at the Intelligent and Mobile Robotics Group of Czech Institute of Informatics, Robotics and Cybernetics, CTU. His main research interests are visual localization, navigation, and computer vision.

Raivo Sell received his Ph.D. degree in Product Development from Tallinn University of Technology in 2007 and currently working as a professor of robotics at TalTech. His research interest covers mobile robotics and self-driving vehicles, smart city, and early design issues of mechatronic system design. He is running the Autonomous Vehicles research group at TalTech as a research group leader with a strong experience and research background in mobile robotics and self-driving vehicles. Raivo Sell has been a visiting researcher at ETH Zürich, Aalto University, and most recently at Florida Polytechnic University in the US, awarded as a Chart Engineer and International Engineering Educator.