Multimodal and multiview deep fusion for autonomous marine navigation

Dimitrios Dagdilelis, Panagiotis Grigoriadis, and Roberto Galeazzi

***Abstract*—This paper introduces a cross-attention transformer-based approach for multimodal sensor fusion, aimed at constructing a bird's-eye view of a vessel's surrounding environment to enhance safe marine autonomous navigation. The proposed method utilizes an innovative model that performs deep fusion of multiview RGB images and long-wave infrared camera images, together with sparse LiDAR point clouds. Additionally, model training incorporates data from X-band radar and electronic nautical charts to generate relevant prediction classes. The reconstructed view offers a detailed representation of the vessel's surroundings, facilitating more accurate and robust navigation. The effectiveness of the proposed approach is evaluated using real-world data collected at sea. Experimental results demonstrate its capability across various scenarios, including challenging weather conditions and complex marine environments.**

***Index Terms*—Sensor fusion, Transformer, Deep fusion, Cross-attention, Bird's-eye view, Multimodal sensor platform, Autonomous marine navigation.**

I. Introduction

Highly automated and autonomous navigation has become a critical area of research in the field of marine robotics due to the growing demand for unmanned maritime operations. However, some of the key challenges in the development of safe autonomous marine systems are to enable a robust and accurate perception of the environment surrounding the vessel and to provide spatial reference for objects in such an environment.

Robust perception depends on the integration of data from diverse sensors, enabling the mitigation of individual sensor limitations through data fusion. A multisensor configuration enhances contextual awareness by leveraging the capabilities of cameras while providing robust detection capabilities—both near and far—via radar and LiDAR that can operate effectively under adverse weather and lighting conditions [1].

Accurate perception requires the determination of the 3D position of other vessels and/or navigation targets [2]. A typical multimodal sensing platform comprises one or more cameras in conjunction with one or several sparse sensors, such as LiDAR or radar. While cameras are limited in their ability to accurately assess 3D geometries due to a lack of depth information, they excel in delivering rich semantic and boundary data. In contrast, although sparse sensors are capable of providing precise short-range 3D positional information, they often fall short in terms of dense semantic content. The integration of dense cameras with sparse sensors represents a promising avenue for advancement. However, existing research in multimodal maritime fusion, although recognizing the importance of this integration, predominantly concentrates on the narrow scope of identifying and tracking foreign vessels [3]–[9].

This emphasis highlights a significant gap compared to the advancements observed in Autonomous Driving (AD) technology, where perception systems have achieved more comprehensive scene understanding capabilities [10]. This broader understanding is crucial for supporting essential autonomous functions, such as collision avoidance and route planning, thereby facilitating effective and safe navigation in diverse environments [10].

The literature review (Sections section II-section III) shows that no previous study has adapted or validated multimodal perception technology specifically for autonomous waterborne navigation. In light of this gap, our research proposes a perception framework designed to process multimodal sensor data from RGB and long-wave infrared (LWIR) cameras, along with LiDAR sensors. This framework aims to facilitate comprehensive scene understanding by predicting, in real-time, the Bird's Eye View (BEV) positions of nearby navigation features. A high-level overview of our approach is illustrated in fig. 1.

The remainder of the paper is structured as follows. Section section II reviews the existing literature on methods for generating BEV models using multi-modal sensor platforms. Section section III formalizes the research problem and reviews existing methods. Section section IV details the development of the proposed approach and discusses how the different sensor technologies are fused through the architecture of the cross-attention transformer. Section section V shows the results of testing the approach using real-world data collected at sea. Sections section VI and section VII draw conclusions and discuss the current limitations of the proposed method and future work.

Table table I provides the list of abbreviations used in the paper.

II. Related Work

Bird's Eye View: BEV recognition models [11]–[15] represent a category of models employed for 3D object detection. These models have garnered both the industry's and academia's research attention, due to their ability to merge partial raw data from heterogeneous perception sensors into a cohesive and comprehensive 3D output space.

D. Dagdilelis, Department of Electrical and Photonics Engineernig, Technical University of Denmark.

P. Grigoriadis, independent contributor.

R. Galeazzi, Department of Electrical and Photonics Engineernig, Technical University of Denmark.

Manuscript received January 1st, 2024; revised January 2nd, 2024.

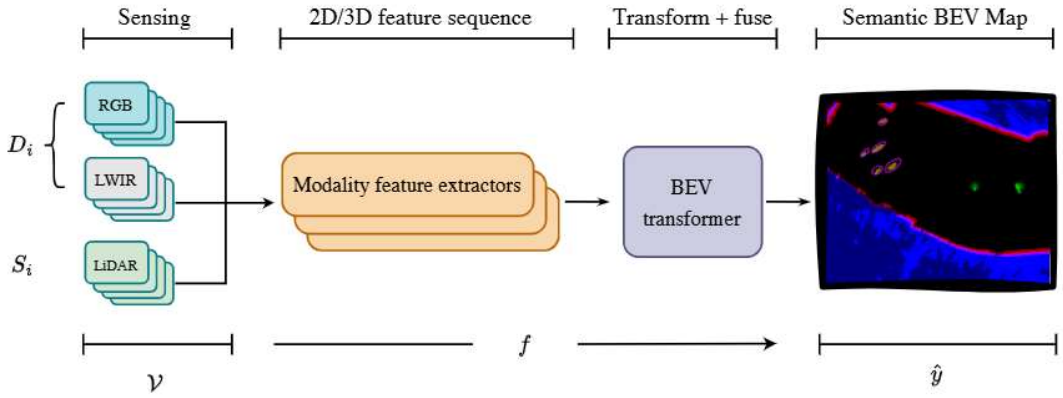

Fig. 1. High level overview of the proposed approach. Sensor streams are processed by modality-specific feature extractors. Modality features are processed as a sequence by the BEV Transformer model into a semantic BEV map. The final output representation is orthographic, compact and semantically useful for downstream navigation tasks.

TABLE I
ABBREVIATIONS

Abbreviation	Explanation
BEV	Bird Eye View
ENC	Electronic Navigational Charts
RCS	Radar Cross Section
IoU	Intersection over Union
XBR	X-band GHz Radar
GNSS	Global Navigation Satellite System
RGB	Red Green Blue
LWIR	Long Wave Infrared

Combining a dense camera with sparse sensors is a promising direction; however, the data collected by these sensors must be mapped to a unified coordinate frame. Naturally, BEV emerges as the ideal representation framework to support such uniform coordinate mapping. More specifically, a BEV representation:

a) is independent of the sensor model and thus easily extendable to additional modalities, b) is geometrically interpretable, c) is the framework that is common to all perception modalities, but at the same time native to downstream navigation tasks, such as motion planning.

According to [16], existing BEV fusion methods are limited because of two reasons. Despite the fact that sparse modalities like LiDAR, can be trivially used to predict 3D features, the same does not apply for camera modalities, where depth information is implicit or missing, and extracting 3D spatial information from monocular or multi-view settings is difficult and largely unsolved [16]. At the same time, many sensor fusion algorithms rely on simple object-level fusion or basic feature concatenation, where misalignment or inaccurate depth predictions between camera and LiDAR data can lead to subpar performance compared to LiDAR-only methods [16].

To overcome these limitations, deep fusion methods have emerged as a promising approach that leverages deep learning models to integrate data from multiple sensors [16]. Among the deep fusion methods, the transformer-based model has shown superior performance in capturing the complex relationships between different modalities and fusing them into a comprehensive representation of the surrounding environment [16].

In this paper, we propose a novel cross-attention transformer-based multimodal sensor fusion approach for marine autonomous navigation. The proposed method integrates information from multiple sensors, including LiDAR, RGB, LWIR, ENC charts, and GPS, and leverages a deep learning model to effectively capture complex relationships between different modalities, in order to predict a semantic BEV map of the surrounding environment. Semantic BEV maps can be used to enable downstream navigation tasks, or be used to perform diagnostics on other sensor modalities such as the maritime radar.

Our experimental results demonstrate the effectiveness of the proposed method in coastal marine navigation environments. Finally, our work contributes to the development of more accurate and robust marine autonomous systems by proposing a deep fusion approach that a) combines sparse and dense perception data, b) integrates the temporal information, c) remains robust to sensor-calibration errors.

III. Problem Formulation

The core idea of BEV perception is to develop a feature representation of information, directed from possibly multiple views or modalities, that can be readily used to solve a downstream task, such as 3D object detection, BEV map segmentation, lane line detection, path planning, or any combination of these tasks. In the following sections, we use *2D* to refer to a perspective view described by camera pixel coordinates, *3D* to refer to the real-world space with world coordinates, and *BEV* to refer to a planar, metric, zero height,

TABLE II
LIST OF KEY MATHEMATICAL NOTATIONS

Symbol	Description
$\mathcal{V}$	Heterogeneous sensor views collection
S	Sparse sensor views collection
D	Dense sensor views collection
$\mathcal{B}$	Multi-class binary BEV maps
N_B	Number of grid coordinates in the BEV map
C_B	Number of semantic BEV map classes
V_i	Number of views in modality i
N_i	Number of elements in modality i
C_i	Number of channels in an element of modality i
t	Number of aggregated data chunk time instances
y	Ground truth BEV map
$\mathcal{G}$	Sparse modality pre-processing method
$S^\star$	Raw sparse point cloud in 3D
S	Post-processsed, dense, pseudo-camera view
$Conv2d$	Convolutional-2D
x^W	World-point in euclidean space
x^I	Image pixel coordinates vector
p_i, p_j	Image pixel coordinates
d	Depth from image plane
r_i	View-aware data feature rays
$f_i^{\tau,k}$	Input sequence feature vector i, corresponding to view k at time instant τ
$\mathcal{Z}$	Set of points in a voxel
$\phi(\cdot)$	Linear layer, embedding position encodings
$c(\cdot)$	Linear layer, embedding camera position vector
$\iota(\cdot)$	Linear layer, embedding the temporal encodings
q_i	Dense ray queries in BEV
b_i	Inverse projection of pixel vectors in the context of camera parameters
m_{BEV}	Latent BEV map feature sequence
Q, K, V	Transformer queries, keys, values
$\rho_i^{\tau,k}$	Direction encoded feature vector
$\tilde{q}_i$	Learnable BEV queries
q_i	Position encoded BEV queries
E_i, I_i	Extrinsic and Intrinsic camera matrices
x^c	Extrinsic and Intrinsic camera matrices
H, W	Input view's height and width
h_q, w_q	Width and height of latent BEV map representation
$n_{\text{BEV}} = h_q \times w_q$	Number of latent BEV map tokens
$\mathcal{U}$	BEV map segmentation decoder

own-ship centered, *Bird's Eye View* world coordinate grid (see figs. 2 and 4).

A. Task definition

The task on which we focus is that of *BEV map segmentation*, i.e., given the features of *multiple* dense or sparse modalities, we are tasked with predicting a *orthographic semantic* BEV map; the task can be formally described as:

Let a heterogeneous collection of multi-modal features $\mathcal{V}$, where

$$\mathcal{V} = \{D_1, \ldots, D_i\} \cup \{S_1, \ldots, S_j\} \tag{1}$$

and

$D_i \in \mathbb{R}^{t \times V_i \times N_i \times C_i}$ and i is a dense 2D modality index.
$S_j \in \mathbb{R}^{t \times V_j \times N_j \times C_i}$ and j is a sparse 3D modality index.
t, V_k, N_k, C_k Represent the number of aggregated time instances, views, elements, and channels of each modality k respectively, with $k \in \{i, j\}$.

Let also a binary multi-class BEV map $y \in \mathcal{B}$, where

$$\mathcal{B} \subseteq \{0,1\}^{N_B \times C_B}$$

with N_B, C_B respectively being the number of elements (map resolution) and classes of the BEV map (see fig. 5).

BEV map segmentation, finds a function f that transforms heterogeneous views $\mathcal{V}$ to binary BEV masks $\hat{y}$, i.e.

$$\hat{y} = f(\mathcal{V}) \quad \text{where} \quad \hat{y} \in \mathcal{B}$$

This problem is particularly complex due to the fact that its inputs and outputs operate within distinct coordinate systems. Structured inputs D_i are captured from calibrated camera views in 2D perspective view, point cloud inputs S_i are captured in 3D, and outputs y are predicted in BEV.

B. Existing methodologies

Existing approaches that model f are characterized depending on: a) *supported modalities*, b) *modality feature extraction mechanics*, c) *2D feature transformation module*. d) *feature fusion mechanics*,

Supported modalities: Camera-only 3D perception is a significant focus in academic research [11], [12], [17]–[25] since it avoids the need of using expensive lidar equipment. The fundamental challenge in camera-only 3D perception, lies in the fact that 2D imaging processes inherently lacks the ability to capture 3D information, making accurate object localization difficult without precise depth estimation.

Lidar only approaches [26]–[33], completely forgo the use of cameras, and hence suffer performing in dense downstream prediction tasks. On the other hand, approaches similar to ours, combine the best of both worlds by fusing camera with lidar features [34]–[38].

Feature extraction: The most widespread approach in multi-modal feature extraction is having individual encoders extracting features from each modality. Concerning image feature extraction, there exists a wealth of 2D perception research, in the form of transfer learning and pre-trained feature extraction backbones [39]. In lidar-only approaches, [26], [28], [32], [40]–[42] use voxelization to structure point-clouds into voxels, followed by a 3D feature extractor, thus retaining data 3D structure but increasing computations. Similarly to our approach, [23], [33], [43]–[48] convert point cloud data into a BEV representation, by discretizing points into a BEV grid. Features such as height, intensity, and density are then extracted from the points within each grid cell to represent the grid's features. However, due to the large number of points in each BEV grid cell, this process can lead to significant information loss.

Feature fusion: Depending on the stage of information fusion we can classify the approaches in two types:

1) Early fusion, refers to combining information at an early stage of processing, such as combining *raw sensor inputs*.

2) Deep fusion, facilitates the interaction of latent modality features within a neural network's structure, leveraging their flexibility to expressing non-linear functions between modalities.

Early fusion in BEV perception operates by decorating one modality with features from the others. In doing so, there is no single optimal decision on which modality is going to be the main carrier of information, as all selections have drawbacks. A widely used early-fusion technique is Painting [34], [49]. Painting projects point cloud data to images to create correspondences and subsequently appends semantic information to the points, while discarding the rest of the information in the images. Point-level decoration with image features is semantically lossy because it suffers from throwing away a lot of contextual information from the cameras due to point-cloud sparsity, which has a severe impact on semantic-oriented tasks like BEV segmentation, while the reciprocal process of image feature decoration is geometrically lossy due to perspective view geometry [38].

Deep fusion leverages the network's capacity to learn complex representations, allowing non-linear interaction between modalities. [20], [50] propose an effective fusion method to transform 2D camera features to BEV features, by efficiently projecting camera features into BEV space and then combining them with lidar BEV features using convolutional layers. In a similar approach, [50], [51] used predicted image depth distributions and common 3D convolutions to generate modality specific voxel spaces that communicate during task prediction, enhancing cross-modal interactions.

Feature transformation Point clouds bear 3D geometry information, and can therefore be transformed in BEV by geometry projection. Camera views on the other hand, lack depth information and transforming them to BEV is non trivial. Existing multi-modal feature fusion methods, are highly dependent on a reliable 2D to BEV image feature transformation. Methodologies for transforming 2D features into BEV representations can be classified into three primary categories based on their approach to depth estimation:

1) Those employing explicit depth prediction.
2) Those utilizing implicit depth prediction.
3) Those that operate without any depth prediction.

This categorization reflects the varying strategies used to address the challenge of projecting planar image features into a three-dimensional space.

Without performing depth prediction, Inverse Perspective Mapping (IPM) [52] introduced a homography derived from the camera's int/extrinsic parameters, and projects from 2D to 3D and vice versa, assuming that the corresponding 3D points lie on a horizontal plane. The basic idea has been used in recent work [53]–[56], with [2], [57] applying for free-space estimation. Violation of the planar assumption and calibration noise create strong artifacts, while performance degrades quickly as distances get longer due lower pixel density at vanishing points and projective geometry (fig. 2).

Lift-splat-shoot (LSS) [24], is a pioneering approach that uses a pre-trained monocular depth prediction model to predict the depth distribution of image features and use it to project them in 3D, addressing the challenging camera-to-BEV transformation problem. Their method resembles pseudo-lidar [58], where depth is used to lift dense pixels into 3d points. Subsequent work in [59] employs a similar approach to LSS to predict categorical depth distribution, but unlike LSS that uses pre-trained depth models, it provides supervision for depth prediction. This line of work is followed by [19], [20], [60]–[62] and extended to stereo vision depth prediction by [61], [63].

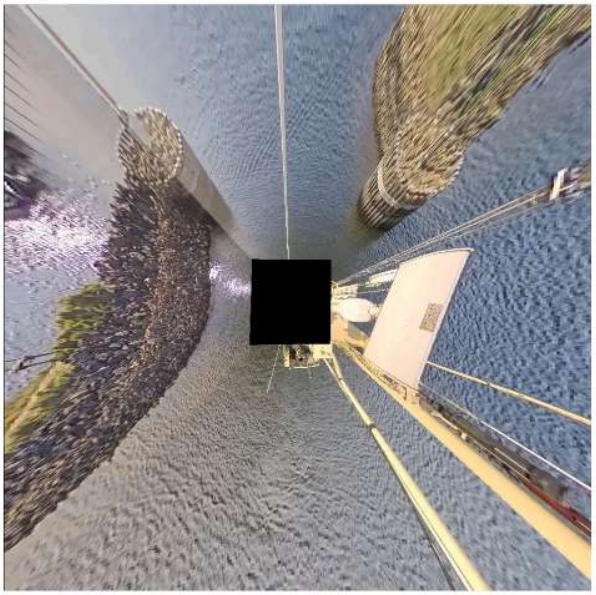

Fig. 2. Outlook from a virtual camera in a top down view pose, based on 360° images from four cameras. The scene is the entrance to a leisure craft harbor at Limfjorden (DK) **??**

Several BEV perception works utilize either multi-layer perceptron [23], [64], [65] or the transformer architecture [66], and implicitly perform depth prediction, in order to transform 2D features to BEV. In a similar way, CVT [13], which greatly influences our work, transforms 2D features to BEV by implicitly predicting depth, and uses learnable camera aware embeddings and the cross attention mechanism to query modality features from BEV coordinates and construct BEV features.

Calibration: Most of the previous work in early and late fusion approaches requires knowledge of precise extrinsic and intrinsic sensor calibration parameters to fuse features between the different coordinate frames, perspectives, and fields of view [57]. Unfortunately, the calibration process can be inaccurate, as illustrated in fig. 6, where world coordinates are projected on image pixels using camera geometry and calibration parameters. The convolution blocks utilized in most existing approaches [38], [67], [68] cannot compensate for dynamic calibration errors, due to their translation invariance. The multiplicity of the sensors, the laborious, time-consuming, and expensive process of calibrating a multi-sensor platform, and the fact that after calibration, the system's performance remains sensitive to external perturbations, have created the need for calibration-free fusion methods. To this end, researchers have explored implicit parameter learning in their models [13],

[69], similar to our approach.

C. Key problems

BEV perception aims to develop a resilient and adaptable feature representation using both camera and LiDAR data. This process is straightforward for LiDAR input, as point clouds inherently contain 3D information. However, it's more challenging for camera input, where extracting 3D spatial data from single or multiple 2D views poses significant difficulties.

Another crucial challenge lies in effectively merging features during the early or middle stages of the processing pipeline. Many existing sensor fusion methods oversimplify this process, either by combining objects at a high level or by naively concatenating features along the data channel. This approach often leads to sub-optimal performance, with some fusion algorithms actually under-performing compared to LiDAR-only solution [16]. The poor performance can be attributed to misalignment issues or inaccurate depth estimates when integrating camera and LiDAR data. Consequently, developing methods to properly align and integrate features from multiple input modalities is critical and presents significant opportunities for innovation in this field.

IV. Proposed Approach

Our proposed method, extends the camera-only CVT [13] by a) appending lidar and LWIR cameras to the operating modalities, b) considering point clouds as pseudo-images and pseudo-views, and processing them in parallel with camera views, c) integrating temporal information from previous timesteps in the fusion process.

To do so, we extract features from individual sensor streams, using modality-specific encoders. Thereafter, we utilize a view transformation module, that utilizes cross-attention guided by position-aware camera encodings, as well as learnable BEV queries, to transform multi-modal sensor features to BEV features. Finally, a decoder upscales the BEV features to the original ground truth BEV map resolution. We provide a high level overview of our approach in fig. 1, while in fig. 11 we provide a more analytical description of the model's components. The model in fig. 11 is end-to-end differentiable, and we optimize it using ground truth BEV maps (read section IV-A) and focal-loss [70].

A. Dataset

Our maritime specific dataset, was collected in Aalborg, Denmark, onboard tugboat *Balder*, during daytime and over a duration of 6 hours. The dataset includes diverse set of modalities, including RGB, LWIR images, two lidar sensors positioned at the front and rear of the vessel and an X-band long-range radar. See table III for a summary of sensor specifications. Additionally, the dataset includes own-ship's geo-location, captured from a satellite compass and a GNSS receiver and static ENC data. Sensors were rigidly mounted on 3m masts (see fig. 3), in order to establish a prominent position and minimize obstructions from vessel structures or other sensors.

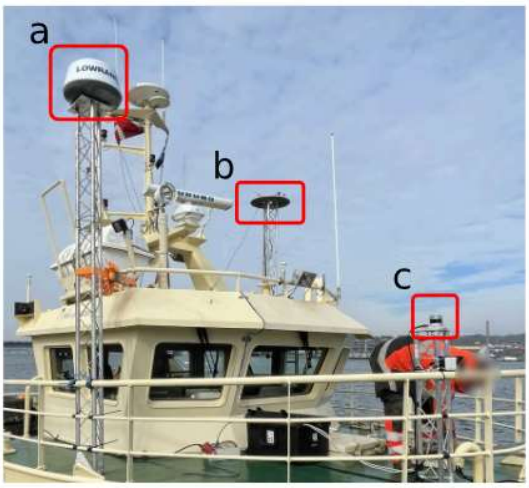

Fig. 3. Tugboat *Balder* used for the data collection at *Limfjorden, DK*. Annotated are: a) XBR b) RGB & LWIR camera platform c) LiDAR

We exploit the complementary nature of the available data sources to construct multi-class BEV ground truth maps $y \in \mathcal{B}$ as illustrated in fig. 4.

ENC are integral to navigation systems and software used on ships and various marine vehicles, playing a critical role in facilitating safe sea travel. The information in them is gathered through surveys and observations conducted by maritime authorities. The information that is relevant in the context of our application, is the location of buoys, land, sail-able water area, and shorelines. We pre-process this information by fusing it with other sensor modalities, generating supervision signals for our model, in the form of BEV maps. To do so, we fuse ENC, with XBR and own-ship geo-location and attitude data, to generate BEV ground truth maps that semantically encode floating buoys, land, shoreline, water volumes and moving targets.

Static classes: The position of objects such as navigational marks, man-made structures, and landscape features, does not change during navigation. The geo-location of such features in the navigation scene, can be trivially extracted from the ENC, requiring no human annotation effort. Our solution aims to take advantage of the following static ENC features:

1) Buoys: Charts detail the precise locations and characteristics of navigational aids, including buoys. Specifically, they provide information about the buoy's type, color, shape, and light characteristics, to assist mariners in identifying and navigating around them.
2) Land areas: Areas occupied by landscape or structures.
3) Shorelines: Outline shape of land areas.
4) Sail-able area: We threshold depth-map contours to extract sail-able water area.

Moving targets: Floating objects within the navigable area are likely to be detected by the XBR, but they may not appear on the ENC. To determine the geographic location of these targets, we use the own-ship's geo-location and attitude measurements to project the XBR data onto navigable water area maps, creating XBR-over-water maps. Since XBR data over water is not affected by ground clutter, it exhibits a high

TABLE III
SENSOR SPECIFICATIONS

Sensor	Max Range	Min Range	Ang. Res.	Range Res.	FoV	Effective FOV
X-band radar	20 NM	85 M	0.225 degrees	1% Max range	360 deg	360
RGB cam	–	–	0.0477 deg	–	(94, 52) deg	360 deg
LWIR cam	–	–	0.078 deg	–	(50, 40) deg	200 deg
Lidar	200m	0.3 m	(0.18, 0.7) deg	12 cm	(360, 45) deg	200 deg
ENC	–	–	–	–	–	360 deg

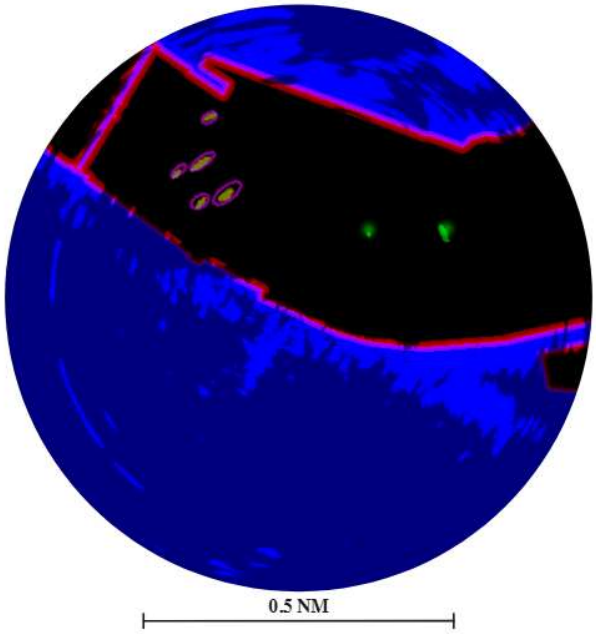

Fig. 4. Ground truth BEV map generation. XBR data are transparently ploted below color-coded ENC classes (blue for land, magenta for shoreline, green for buoy, black for water). We curate *moving target* instances, by annotating magenta ellipsoids on top of verified moving targets.

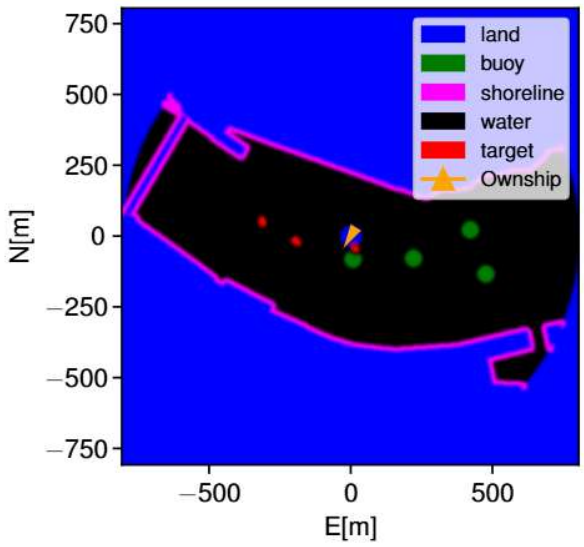

Fig. 5. Creating a ground truth BEV map y using XBR, ENC, satellite-compass, and GNSS.

signal-to-noise ratio, making it probable that echoes detected in these XBR-over-water maps correspond to actual moving targets. We then manually review the identified blobs in the XBR-over-water maps, cross-referencing them with camera data (see fig. 6) to confirm the presence of moving targets and eliminate false positives. This process allows us to accurately determine the geographic location of moving targets.

High-frequency features and XBR noise are filtered by applying a series of dilation erosion and Gaussian smoothing. The derived XBR-over-water maps, are manually annotated, using camera images as cross-referencing data, marking the BEV position of the *targets* class (see fig. 4) with ellipsoids.

The orthographic property of BEV maps, enables a direct and consistent relationship between pixel coordinates in the BEV map and local North-East-Down coordinates. While not strictly correct, the assumption of all BEV features residing on the same plane is valid in our operation environment. Consequently, the position of any instance or object within a BEV map can be converted to geodetic coordinates (latitude, longitude, and altitude) through the application of a local geodetic model. The process involves an initial conversion from pixel coordinates to local Cartesian coordinates, followed by a transformation to geodetic coordinates using the specified local geodetic framework.

B. Solution architecture

The key components of our solution architecture are a) a pseudo-camera view pre-processing method G that rasterizes sparse and unstructured point clouds to structured and dense virtual camera views, b) the modality specific feature encoders, c) the positional-ray encodings, d) a transform and fuse module that generates BEV features from both actual and virtual camera views, e) a decoder that upscales the BEV features to the final BEV map shape.

LiDAR preprocessing:

Raw point-clouds are unstructured and sparse, and hence not compatible with most deep learning processing blocks, that expect structured vectors of defined shapes. Inspired by [33], [44], [71], we pre-process LiDAR point clouds and conver them to pseudo-images, enabling their processing with convolution architectures.

Let $S^\star$ represent a 3D point cloud with $N_{S^\star}$ points

$$S^\star = \{(x_n, y_n, z_n) : i \in \mathbb{N}_0, n \leq N_{S^\star}\}$$

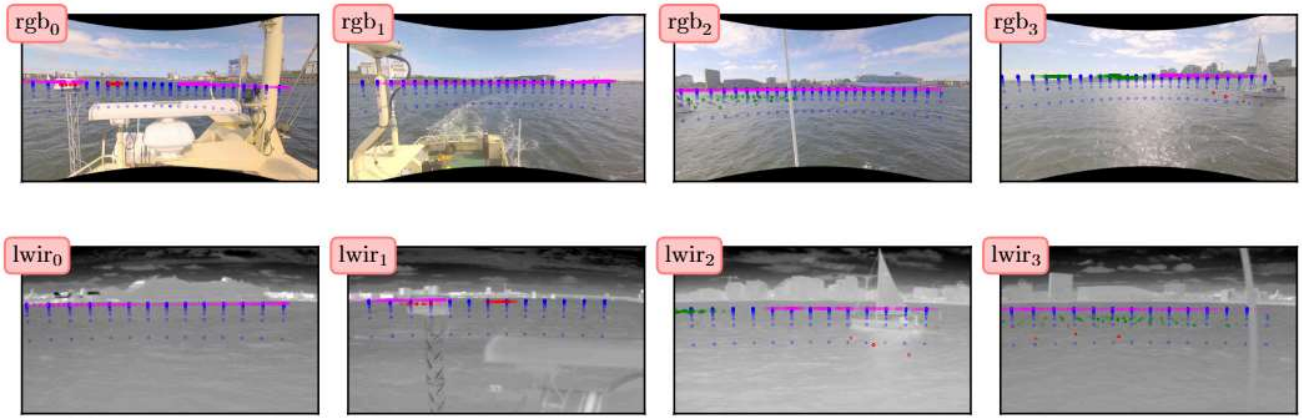

Fig. 6. Camera view corresponding to fig. 5. The semantic map y in fig. 5 is sparsely projected on the camera views (note the color inversion between blue and black). Projection errors are visible in camera views, due to inaccurate calibration parameters or noise in the own-ship pose.

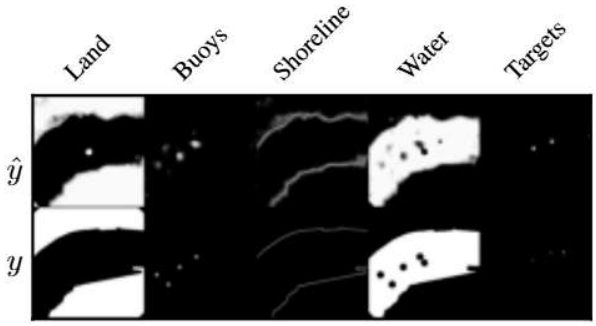

Fig. 7. Softmaxed BEV map predictions $\hat{y}$. Each BEV map in the figure's tiles, predicts the presence of one of the 5 classes of interest, within 600x600m area around own-ship, where pixel intensity encodes the class's likelihood.

Let also S represent the processed pseudo-image of $S^\star$

$$S = \mathcal{G}(S^\star) \in \mathbb{R}^{H \times W \times C_S}$$

where

$\mathcal{G}$: $\mathbb{R}^3 \to \mathbb{R}^{N_{\text{LiDAR}} \times C_{\text{LiDAR}}}$
S: $\{f_{ij} : i \in \{0, 1, \dots, H\},\ j \in \{0, 1, \dots, W\}\}$
f_{ij}: $\left(\mathbb{E}\left[\mathcal{Z}_{ij}\right] \quad \text{Var}\left(\mathcal{Z}_{ij}\right) \quad \max(\mathcal{Z}_{ij}) \quad \min(\mathcal{Z}_{ij})\right)^T$
$\mathcal{Z}_{ij}$: Aggregated points within each voxel given by eq. (2)
$\lfloor \cdot \rfloor$: Rounds down to the closest integer
H, W: Pseudo-image dimensions, such that $HW = N_{\text{LiDAR}}$ in eq. (1) notation
d_S: Spatial dimension of the pseudo-image grid (see fig. 10)

$$\mathcal{Z}_{ij} = \{z_n | (x_n, y_n, z_n) \in S^\star | \left\lfloor \frac{x_n}{d_S} \right\rfloor = i \wedge \left\lfloor \frac{y_n}{d_S} \right\rfloor = j\} \quad (2)$$

In our application, the sparse modality corresponds to the LiDAR sensors, therefore we introduce the $\{\cdot\}_{\text{LiDAR}}$ subscript to ground our notation. We preprocess and standardize the unstructured and irregularly sized point clouds $S^\star_{\text{LiDAR},u,\tau}$ to their structured and uniform BEV pseudo-image representations $S_{\text{LiDAR},u,\tau}$, while preserving 3D geometry features

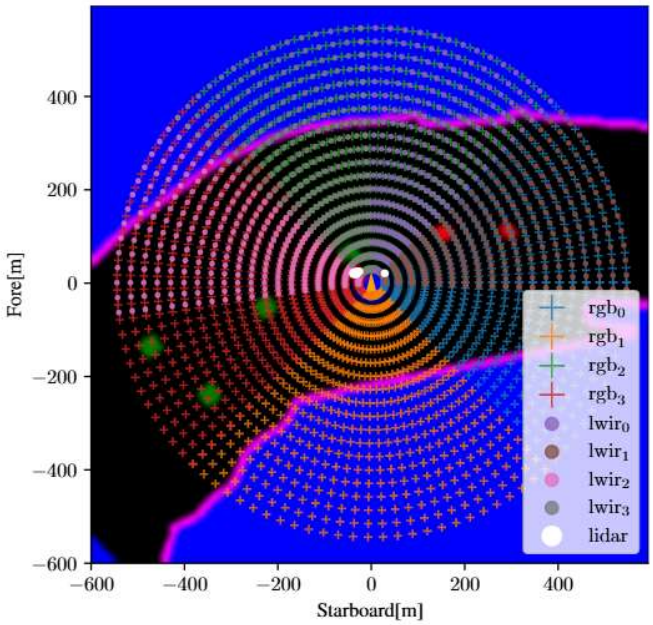

Fig. 8. Individual modality field of views annotated by maker type and color, projected in BEV, illustrating field of view overlap.

through points statistics. This processing is illustrated in fig. 10. Finally, we aggregate processed pseudo-images for all V_{LiDAR} lidar sensors and time instances t, to form $S_{\text{LiDAR}} \in \mathbb{R}^{t \times V_{\text{LiDAR}} \times N_{\text{LiDAR}} \times C_{\text{LiDAR}}}$, matching eq. (1) notation.

$$S_{\text{LiDAR}} = \{\mathcal{G}(S^\star_{\text{LiDAR},v,\tau}) : v, \tau \in \mathbb{N} \mid v \leq V_{\text{LiDAR}},\ \tau \leq t\}$$

Modality encoders:

We revisit eq. (1), by introducing further grounding subscript notation to the dense modalities.

$$\mathcal{V} = \{D_{\text{RGB}}, D_{\text{LWIR}}\} \cup \{S_{\text{LiDAR}}\}$$

- $D_{\text{RGB}}, D_{\text{LWIR}} \in \mathbb{R}^{t \times \mathcal{V}_D \times H \times W \times 3}$ represent dense monocular camera views,
- $S_{\text{LiDAR}} \in \mathbb{R}^{t \times V_{\text{LiDAR}} \times N_{\text{LiDAR}} \times C_{\text{LiDAR}}}$ represents the raw and sparse 3D point-cloud of lidar v at time instant τ.

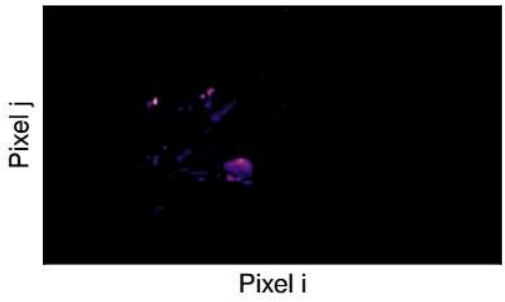

Fig. 9. PCA (n=1) of a pseudo-camera lidar point cloud S (see also fig. 10).

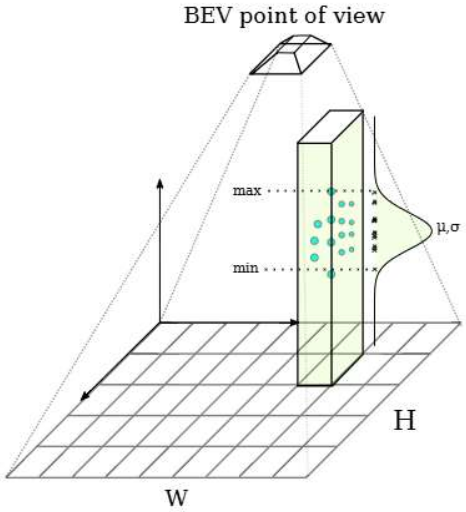

Fig. 10. Pre-processing $\mathcal{G}$ of LiDAR point cloud data. The points are projected on a BEV grid defined by the view a BEV point of view and the $z = 0$ ground place. Each cell contains statistics of the z-coordinate values of the points that fall within the cell (see also fig. 9).

We encode modalities into latent features, using modality-specific backbones at different resolutions. In order to extract features from RGB and LWIR camera views, we convert LWIR images to RGB, and extract feature maps from the intermediate layers of a shared pre-trained EfficientNet-b4. Since we can't use transfer learning to extract features from the LiDAR pseudo-views $S_i^\star$, we use a randomly initialized U-Net [72] feature extractor, similarly to [73]. Modality encoders are projecting their features to a common dimension, hence we adopt the notation feature $f_i^{\tau,k} \in \mathbb{R}^{d_e}$ as the feature extracted from camera or pseudo-camera k at time index τ.

Cross attention For any world point $x^W \in \mathbb{R}^3$, there exists a corresponding image pixel coordinate $x^I = (p_i, p_j, d)$. For a view k, world points are connected to camera coordinates by eq. (3):

$$x^I = I_k E_k (x^W - x^c{}_k) \tag{3}$$

Where $I_k, E_k \in \mathbb{R}^{3\times3}, x^c{}_k \in \mathbb{R}^3$ represent the int/extrinsic matrices and positioning vector respectively, of view k. Without knowledge of depth from the image plane d, the per-point inverse projection in eq. (3) is under-defined. We argue that depth information is implicitly represented in a scene, and transformer models can learn to create BEV maps, by querying for BEV elements in a key, value sequence, consisting of input features across all views and time instances (see fig. 12). In order to help the model develop BEV-to-view feature correspondences, we suggest using a view's int/extrinsic parameters, to positionally encode an input feature $f_i^{\tau,k}$ with its view-aware direction embedding, and effectively guide the positional encoding process with camera geometry priors. We do this by creating unit direction vectors $\rho_i^{\tau,k} = {E_k}^{-1}{R_k}^{-1}{x^I}_i^{\tau,k}\big|_{d=1}$ for every feature $f_i^{\tau,k}$ and its view k pixel coordinates ${x^I}_i^{\tau,k}$ at time index τ.

Instead of using per-feature back-projection to generate BEV features, and following the success of [11], [13], [14] we use the function of cross-attention in the transformer decoder [66] to express the problem as a sequence-to-sequence translation [74].

$$\overbrace{m_{BEV}}^{\text{output sequence}} = \mathcal{X}(\overbrace{\{0,1,\ldots,q_{\text{BEV}}\}}^{Q}, \overbrace{\{0,1,\ldots,r_i^{\tau,k}\}}^{K}, \overbrace{\{0,1,\ldots,f_i^{\tau,k}\}}^{V}) \tag{4}$$

where the output and query sequence $m_{BEV}, Q \in \mathbb{R}^{n_{\text{BEV}}\times d_m}$ consist of map feature tokens and learnable query tokens, the input sequence consists of features $f_i^{\tau,k} \in \mathbb{R}^{d_e}$. The output sequence is obtained by quering the direction encoded input sequence with learnable map-queries Q.

$$r_i^{\tau,k} = f_i^{\tau,k} + \phi(\rho_i^{\tau,k}) + \iota(\tau)$$

$$q_i = \tilde{q}_i + \phi(b_i) - \epsilon({x^c}_k)$$

$$b_i = E_q^{-1} I_q^{-1} \left[b_{x_i}, b_{y_i}, 1\right]^T, i \in \{0,1,\ldots,h_q\}\times\{0,1,\ldots,w_q\}$$

where:

- $r_i^{\tau,k}$ are geometry aware feature rays,
- I_q, E_q int/extrinsic matrices of a pseudo-camera observing the BEV map, i.e. a pseudocamera with focal length 1, positioned at $(0,0,1)$ and directed towards $-z$,
- $\phi, \epsilon : \mathbb{R}^3 \to \mathbb{R}^{d_e}$, $\iota : \mathbb{R} \to \mathbb{R}^{d_e}$ are linear projection layers, making sure the dimension of the position embeddings match the dimension of the features,
- f_i corresponds to the feature vector of pixel x_i^I extracted from the associated backbone,
- $q_i \in \mathbb{R}^{d_m}$ are learnable map-queries,
- d_m, d_e are the latent map feature dimensions, and the input feature dimensions respectively.

Decoder:

The semantic segmentation decoder $\mathcal{U}$, consists of three bilinear upsampling layers, that scale the m_{BEV} features, to match the latent representation of y, i.e.

$$\hat{y} = \mathcal{U}(m_{\text{BEV}})$$

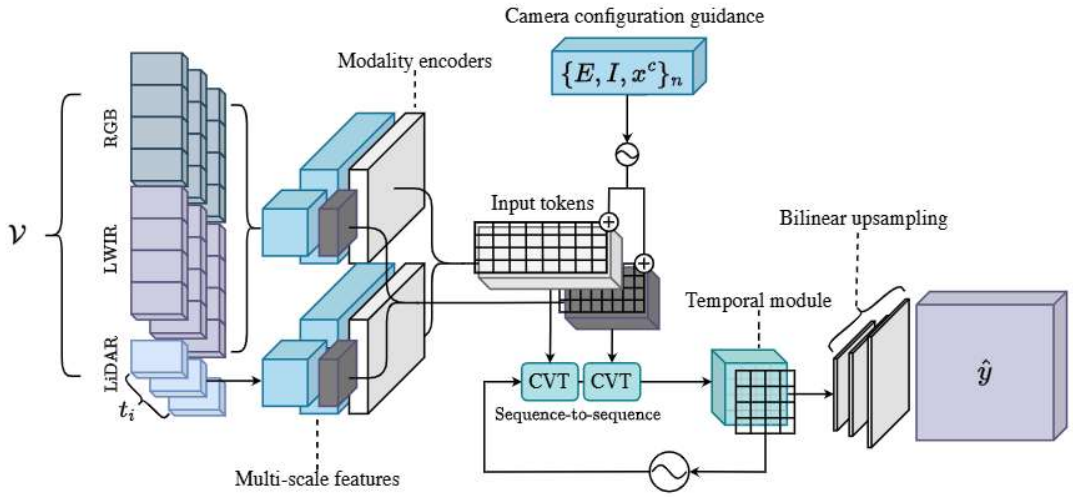

Fig. 11. Our proposed fusion framework: We present a cross-modal and cross-view fusion architecture for the task of BEV segmentation based on camera images and (LiDAR) point clouds. The dense map-queries allow the model to attend to specific regions within each modality, and the multi-scale implementation allows capturing small and large scale context. Temporal information is encoded by 3d convolution blocks.

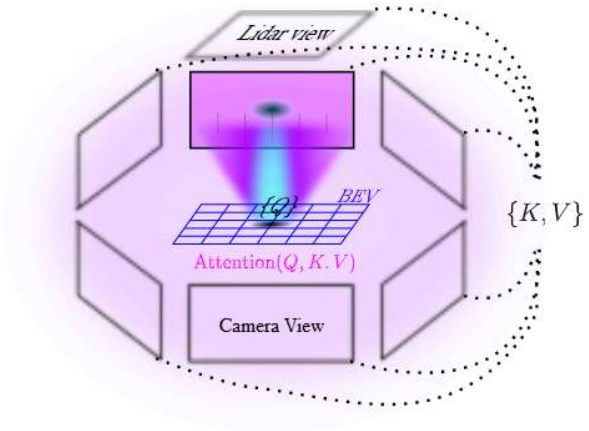

Fig. 12. View-aware cross attention. Pixel positions are lifted to unit direction vector coordinates, which are appended to the features as positional embeddings.

C. Motivation for Multi-modality

The principle of multi-modality posits that the integration of additional sensors, in conjunction with traditional RGB cameras, can potentially enhance the model's ability to reason the depth of the ship's surroundings and the category to which they belong. Regarding depth, the infrared and point cloud data inherently contain that information. Focusing on the segmentation aspect, the lidar and radar data provide localization information that can not be found in camera images. As a result, we hypothetize that downstream task performance can be improved given their effective deep fusion.

D. Temporal features

By integrating historical BEV information with current environmental data, models can stabilize perception results, represent temporarily occluded objects, accumulate observations for map generation, and improve the detection of moving or heavily occluded objects [20], [75]. In our work, four out of five classes are static (buoys, land, shoreline, water). The displacement of these classes between consecutive images and BEV maps therefore, depends exclusively on the own-ship's motion. The ship's pose is, in most cases, measurable through GNSS.

The *target* class, however, might display non-linear motion dynamics, and the displacement of this class's instances between consecutive images and BEV maps, depends on the combined relative displacement between the moving targets and own-ship. The *target* instances motion dynamics are therefore relatively unpredictable, especially without being able to measure their velocity. We hypothesize that by allowing a model to simultaneously process temporal information across multiple time instances, it can learn object motion dynamics and avoid occlusion, as demonstrated in [11].

We extend the work of [13] and incorporate a temporal aggregation module, acting on heterogeneous multi-view features. We attempt this by using ego-motion to align the output of the cross-attention layers across different time instants, and then extracting spatio-temporal BEV features, with 3D convolution kernels.

Temporal Alignment To align the BEV feature maps from consecutive time instances, similarly to [19], [22], [76], we use ego-motion pose differences, to align the feature maps between past/future time instances.

3D Convolutional Layers The output of the alignment process are t aligned BEV feature maps. We then use a 3D convolution layer without padding, to extract spatio-temporal BEV features.

V. Experiments

A. Implementation details

Model parameters: We use transfer learning and EfficientNet [77] checkpoints, to implement the common RGB and LWIR camera feature extractor. For the LWIR images, which are single-channel, we employ channel replication to create a three-channel input compatible with the RGB-based network architecture. We standardize all images to 224x480 resolution and similar to [13], we extract features in two 8x and 16x down-scaled resolutions, i.e., (28,60) and (14,30) and $d_e = 128$ channels.

For the learnable BEV queries, we choose $n_{\text{BEV}} = 25 \times 25 = 625$ and $d_m = 128$. For the multi-head cross-attention blocks, we choose 4 heads and inner dimension 64. The decoder consist of three bi-linear upsampling and convolution layers, upscaling the latent map feature tokens m_{BEV} to the prediction outputs $\hat{y} \in \mathbb{R}^{200\times200\times5}$, which corresponds to 3 m/pixel orthographic projection maps, one for each class, covering 600m in width and height (see also fig. 7).

For the temporal aggregation module, we choose $t = 3$ time instances in eq. (1) and select samples with 5 s difference between them.

Augmentations: BEV Perception needs to remain robust to geometrical perturbations in images. Such perturbations are either non-constant, for example as consequence of the vessel's floating movements, or static, i.e. due to imprecise calibration parameters. The goal of train-time augmentations is to make the model invariant to such perturbations, but also regularize the training process, by reducing over-fitting. Train-time data augmentation plays a vital role in improving BEV mapping performance, with [78] reporting substantial benefits of train-time augmentations in their ablation studies.

We apply random geometric augmentation to the sensor views, implemented with random cropping and rotation, where we update the int/extrinsic matrices correspondingly [79]. Following the geometric transformations, we apply random color variation augmentations, trying to cover different light exposure conditions, as well as random camera dropout augmentation, where one or more of the sensors are being randomly blacked out during training.

Training: We train on 2 NVidia A100 GPUS for 120 epochs, over 8 hours with batch size of 12 samples, using binary focal loss. We optimize using AdamW [80] optimizer with cyclical learning rate scheduling [81] and a maximum learning rate of $4 \times 10^{-}3$.

Evaluation: We apply softmax and argmax to the predicted BEV maps $\hat{y}$ across the channel dimension and calculate the multi-class Intersection Over Union (IOU). We repeat the evaluation over a multitude of experiments, and verify the performance benefits of modal and temporal fusion. The evaluation is summarized in table IV. Furthermore, we empirically evaluate the predictive performance of the model by looking at individual class prediction heat-maps in fig. 7, but also by back-tracing salience heta-maps from the cross-attention similarity matrices in the transformer blocks (fig. 14).

B. Explainability

We use salience maps, to demonstrate the model's ability to focus on specific regions within the views $\mathcal{V}$. Figure 14 illustrates the use of extracted attention maps within the model's cross attention blocks to generate saliency maps and overlay them on the camera views. The illustration demonstrates that the model is able to attest to significant parts of the input sequence $\mathcal{V}$ during task inference. The attention mechanism allows the model to assign different weights to different parts of the data, providing insight into which areas influence the model's predictions the most. By revisiting eq. (4), the look-up function is implemented by

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathbf{W}\mathbf{V}$$

$$\mathbf{W} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \in \mathbb{R}^{n_h \times (tVN) \times N_B}$$

where n_h represents the number of heads in Multi-head attention.

To visualise the attention heat-maps, we select a region of interest in the produced BEV maps, that corresponds to a set of indices $n_{\text{probe}} < N_B$ such that $\mathbf{W}_{\text{probe}} \subset \mathbf{W}$. The attention matrix $\mathbf{W}_{\text{probe}}$ is indexed on current time instance, while information along the number of heads n_h is reduced by projecting $\mathbf{W}$ on its first principal component along the head dimension. The attention maps are bi-linearly up-sampled to match the input camera view resolution, resulting in attention matrices $\mathbf{W}^v_{\text{probe}} \in \mathbb{R}^{H\times W}$, for $v = \{1, \ldots, V\}$ that can be visualised on top camera views, as in fig. 14.

VI. Conclusion

We present a transformer based, calibration-robust map-view segmentation approach, that integrates RGB, LWIR and LiDAR sensors across multiple time instances. We validate our approach on field collected, maritime specific data, demonstrating remarkable potential in long-range scene understanding and suggesting that our that BEV perception has a well-functional position in the design of autonomous ship-borne navigation systems.

VII. Limitations and Future Work

Spatio-temporal features Despite the positive performance boost (table IV) of using 3D convolution and ego-motion alignment in our suggested temporal aggregation module, our approach is still dependant on accurate ego-motion estimation information. At the same time, convolutions have local receptive fields, and are thus unable to model long spatio-temporal dependencies. Further work is required in quantifying the effect of spatio-temporal aggregation in BEV perception.

Doppler radar: Naturally, properties such as the radial velocity in doppler velocity radars, exhibit homogeneity for points originating from a single object. Therefore, it is reasonable to hypothesize that integrating such information will be beneficial in the BEV map segmentation task. In our experiments, the W-Radar modality was excluded due to its sparse sampling, which made it unsuitable for use. However,

TABLE IV
RESULTS IN TERMS OF MEAN INTERSECTION OVER UNION (MIOU).

Method	Modality	Boat	Buoy	Water	Shoreline	Land
Standard	RGB	11%	30%	70%	22%	92%
Standard	RGB, LWIR	13%	32%	78%	23%	93%
Standard	RGB, LWIR, LiDAR	15%	34%	80%	26%	95%
Temporal	RGB, LWIR, LiDAR	17%	36%	85%	27%	96%
Temporal w. alignment	RGB, LWIR, LiDAR	17%	37%	91%	31%	96%

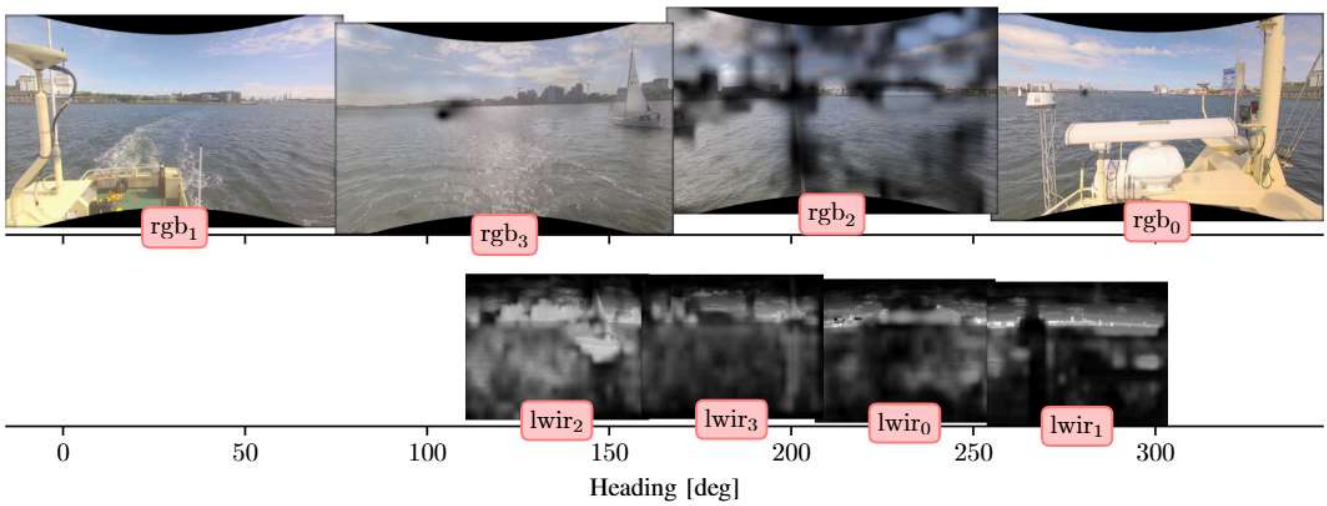

Fig. 13. Overlaying attention map on the input data. The figure illustrates the correlation between BEV and camera view representations. Image position on each row adjusted for camera k extrinsic parameters I_k (see also fig. 14).

as recently highlighted by [73], [79], BEV perception could benefit from further investigation into the fusion of the W-Radar modality, suggesting it as a promising area for future research. It is worth mentioning that our approach readily supports the integration of W-Radar data as an additional modality, by aggregating them into additional pseudo-views.

Cross-attention: While capable of capturing global correspondences across different map locations and views, the complexity of cross attention scales linearly with the number resolution of the BEV map, and quadratically with input feature resolution and number of timesteps, and number of views, prohibiting the use of fine-detailed feature maps or BEV maps. At the same time, a lot of calculations are possibly redundant as global cross attention, naively calculates scores between BEV map locations and views, that do not necessarily correspond to each other. The use of deformable attention [82], as already demonstrated in [11], [78], [79] is a promising direction in reducing the computation complexity of cross attention, and at the same time enables the selective attendance of a BEV query to specific views regions.

ACKNOWLEDGMENT

This research was sponsored by the Danish Innovation Fund, The Danish Maritime Fund, the Orients Fund, and the Lauritzen Foundation through the Autonomy part of the ShippingLab project, grant number 8090-00063B. The electronic navigational charts were provided by the Danish Geodata Agency.

REFERENCES

[1] A. Singh, "Vision-RADAR fusion for Robotics BEV Detections: A Survey," 2 2023.

[2] M. T. Paasche, O. K. Helgesen, and E. F. Brekke, "Real-time 360 degrees view for the operator of milliampere 2," *Journal of Physics: Conference Series*, vol. 2618, no. 1, 2023.

[3] F. Farahnakian, M. H. Haghbayan, J. Poikonen, M. Laurinen, P. Nevalainen, and J. Heikkonen, "Object detection based on multi-sensor proposal fusion in maritime environment," *Proceedings - 17th Ieee International Conference on Machine Learning and Applications, Icmla 2018*, p. 8614183, 2018.

[4] R. Ma, Y. Yin, and K. Bao, "Ship detection based on lidar and visual information fusion," *2022 Conference on Lasers and Electro-optics, Cleo 2022 - Proceedings*, p. JW3B.12, 2022.

[5] Z. Yao, X. Chen, N. Xu, N. Gao, and M. Ge, "Lidar-based simultaneous multi-object tracking and static mapping in nearshore scenario," *Ocean Engineering*, vol. 272, p. 113939, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801823003232

[6] J. Han, Y. Cho, J. Kim, and P. Lee, "Comparison of multi-sensor fusion methods for maritime target object tracking," *Journal of Institute of Control, Robotics and Systems*, vol. 25, no. 6, pp. 551–556, 2019.

[7] M. H. Haghbayan, F. Farahnakian, J. Poikonen, M. Laurinen, P. Nevalainen, J. Plosila, and J. Heikkonen, "An efficient multi-sensor fusion approach for object detection in maritime environments," *Ieee Conference on Intelligent Transportation Systems, Proceedings, Itsc*, vol. 2018-, pp. 2163–2170, 2018.

[8] R. Douguet, D. Heller, and J. Laurent, "Multimodal perception for obstacle detection for flying boats-unmanned surface vehicle (usv)," in *OCEANS 2023-Limerick*. IEEE, 2023, pp. 1–8.

[9] O. K. Helgesen, E. F. Brekke, H. H. Helgesen, and O. Engelhardtsen, "Sensor combinations in heterogeneous multi-sensor fusion for maritime target tracking," *Fusion 2019 - 22nd International Conference on Information Fusion*, p. 9011297, 2019.

[10] T. A. Nygard, N. Dalhaug, R. Mester, E. Brekke, and A. Stahl, "Stereo camera-based free space estimation for docking in urban waters," *Modeling Identification and Control*, 2024.

[11] 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.

[12] 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.

[13] B. Zhou and P. Krähenbühl, "Cross-view Transformers for real-time Map-view Semantic Segmentation," 5 2022.

[14] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view Semantic Segmentation for Sensing Surroundings," 6 2019.

[15] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic Feature Transform for Monocular 3D Object Detection," 11 2018.

[16] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, J. Zeng, Z. Li, J. Yang, H. Deng, H. Tian, E. Xie, J. Xie, L. Chen, T. Li, Y. Li, Y. Gao, X. Jia, S. Liu, J. Shi, D. Lin, and Y. Qiao, "Delving into the Devils of Bird's-eye-view Perception: A Review, Evaluation and Recipe," Sep. 2023.

[17] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 913–922.

[18] Z. Liu, Z. Wu, and R. Tóth, "Smoke: Single-stage monocular 3d object detection via keypoint estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 996–997.

[19] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[20] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[21] L. Reiher, B. Lampe, and L. Eckstein, "A sim2real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–7.

[22] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "Fiery: Future instance prediction in bird's-eye view from surround monocular cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 273–15 282.

[23] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "Hdmapnet: An online hd map construction and evaluation framework," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4628–4634.

[24] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.

[25] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 533–549.

[26] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 529–10 538.

[27] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[28] L. Fan, Z. Pang, T. Zhang, Y.-X. Wang, H. Zhao, F. Wang, N. Wang, and Z. Zhang, "Embracing single stride 3d object detector with sparse transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8458–8468.

[29] Y. Hu, Z. Ding, R. Ge, W. Shao, L. Huang, K. Li, and Q. Liu, "Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 969–979.

[30] J. Mao, M. Niu, H. Bai, X. Liang, H. Xu, and C. Xu, "Pyramid r-cnn: Towards better performance and adaptability for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2723–2732.

[31] X. Chen, S. Shi, B. Zhu, K. C. Cheung, H. Xu, and H. Li, "Mppnet: Multi-frame feature intertwining with proxy points for 3d temporal object detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 680–697.

[32] 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.

[33] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[34] 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.

[35] T. Yin, X. Zhou, and P. Krähenbühl, "Multimodal virtual point 3d detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 494–16 507, 2021.

[36] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, F. Zhao, B. Zhou, and H. Zhao, "Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection," *arXiv preprint arXiv:2201.06493*, 2022.

[37] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le *et al.*, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 182–17 191.

[38] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation," 5 2022.

[39] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4918–4927.

[40] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[41] C. He, H. Zeng, J. Huang, X.-S. Hua, and L. Zhang, "Structure aware single-stage 3d object detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 873–11 882.

[42] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu, "Voxel transformer for 3d object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3164–3173.

[43] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-View 3D Object Detection Network for Autonomous Driving," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 6526–6534, 11 2016.

[44] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7652–7660.

[45] B. Yang, M. Liang, and R. Urtasun, "Hdnet: Exploiting hd maps for 3d object detection," in *Conference on Robot Learning*. PMLR, 2018, pp. 146–155.

[46] Y. Zeng, Y. Hu, S. Liu, J. Ye, Y. Han, X. Li, and N. Sun, "Rt3d: Real-time 3-d vehicle detection in lidar point cloud for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3434–3440, 2018.

[47] W. Ali, S. Abdelkarim, M. Zidan, M. Zahran, and A. El Sallab, "Yolo3d: End-to-end real-time 3d oriented object bounding box detection from lidar point cloud," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[48] M. Simony, S. Milzy, K. Amendey, and H.-M. Gross, "Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds," in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[49] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 11 794–11 803.

[50] 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," 2022.

[51] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," 6 2022.

[52] H. A. Mallot, H. H. Bülthoff, J. J. Little, and S. Bohrer, "Inverse perspective mapping simplifies optical flow computation and obstacle detection," *Biological cybernetics*, vol. 64, no. 3, pp. 177–185, 1991.

[53] N. Garnett, R. Cohen, T. Pe'er, R. Lahav, and D. Levi, "3D-LaneNet: End-to-End 3D Multiple Lane Detection." [Online]. Available: https://sites.google.com/view/danlevi/3dlanes

[54] Y. Kim and D. Kum, "Deep learning based vehicle position and orientation estimation via inverse perspective mapping image," *IEEE*

Intelligent Vehicles Symposium, Proceedings, vol. 2019-June, pp. 317–323, 6 2019.

[55] A. Loukkal, Y. Grandvalet, T. Drummond, and Y. Li, "Driving among Flatmobiles: Bird-Eye-View occupancy grids from a monocular camera for holistic trajectory planning."

[56] J. Philion and S. Fidler, "Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D," 8 2020.

[57] M. K. Plenge-Feidenhans'l, "Robust free area mapping for autonomous harbour navigation," Ph.D. dissertation, 2023.

[58] Y. Wang, W. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," *CoRR*, vol. abs/1812.07179, 2018.

[59] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," *CoRR*, vol. abs/2103.01100, 2021.

[60] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485.

[61] X. Guo, S. Shi, X. Wang, and H. Li, "Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector," *2021 Ieee/cvf International Conference on Computer Vision (iccv)*, pp. 3133–3143, 2021.

[62] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, "Is pseudo lidar needed for monocular 3d object detection," in *2021 IEEE CVF International Conference on Computer Vision ICCV*, 2021, pp. 3122–3132.

[63] Y. Chen, S. Liu, X. Shen, and J. Jia, "Dsgn: Deep stereo geometry network for 3d object detection," *2020 Ieee/cvf Conference on Computer Vision and Pattern Recognition (cvpr)*, pp. 12 533–12 542, 2020.

[64] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *2022 International conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 9200–9206.

[65] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

[66] 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.

[67] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, "Simple-BEV: What Really Matters for Multi-Sensor BEV Perception?" 6 2022.

[68] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "HDMapNet: An Online HD Map Construction and Evaluation Framework," 7 2021.

[69] Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position Embedding Transformation for Multi-View 3D Object Detection," 3 2022.

[70] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[71] J. Beltrán, C. Guindel, F. M. Moreno, D. Cruzado, F. Garcia, and A. De La Escalera, "Birdnet: a 3d object detection framework from lidar information," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3517–3523.

[72] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[73] N. Hendy, C. Sloan, F. Tian, P. Duan, N. Charchut, Y. Xie, C. Wang, and J. Philbin, "Fishing net: Future inference of semantic heatmaps in grids," 2020.

[74] C. Yun, S. Bhojanapalli, A. S. Rawat, S. J. Reddi, and S. Kumar, "Are Transformers universal approximators of sequence-to-sequence functions?" *8th International Conference on Learning Representations, ICLR 2020*, 12 2019.

[75] Z. Qin, J. Chen, C. Chen, X. Chen, and X. Li, "Unifusion: unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view," 2022.

[76] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "BEVerse: Unified Perception and Prediction in Birds-Eye-View for Vision-Centric Autonomous Driving," 5 2022.

[77] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[78] C. Yang, Y. Chen, H. Tian, C. Tao, X. Zhu, Z. Zhang, G. Huang, H. Li, Y. Qiao, L. Lu *et al.*, "Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 830–17 839.

[79] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, "Simple-bev: What really matters for multi-sensor bev perception?" in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2759–2765.

[80] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[81] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006. SPIE, 2019, pp. 369–386.

[82] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," 10 2020.

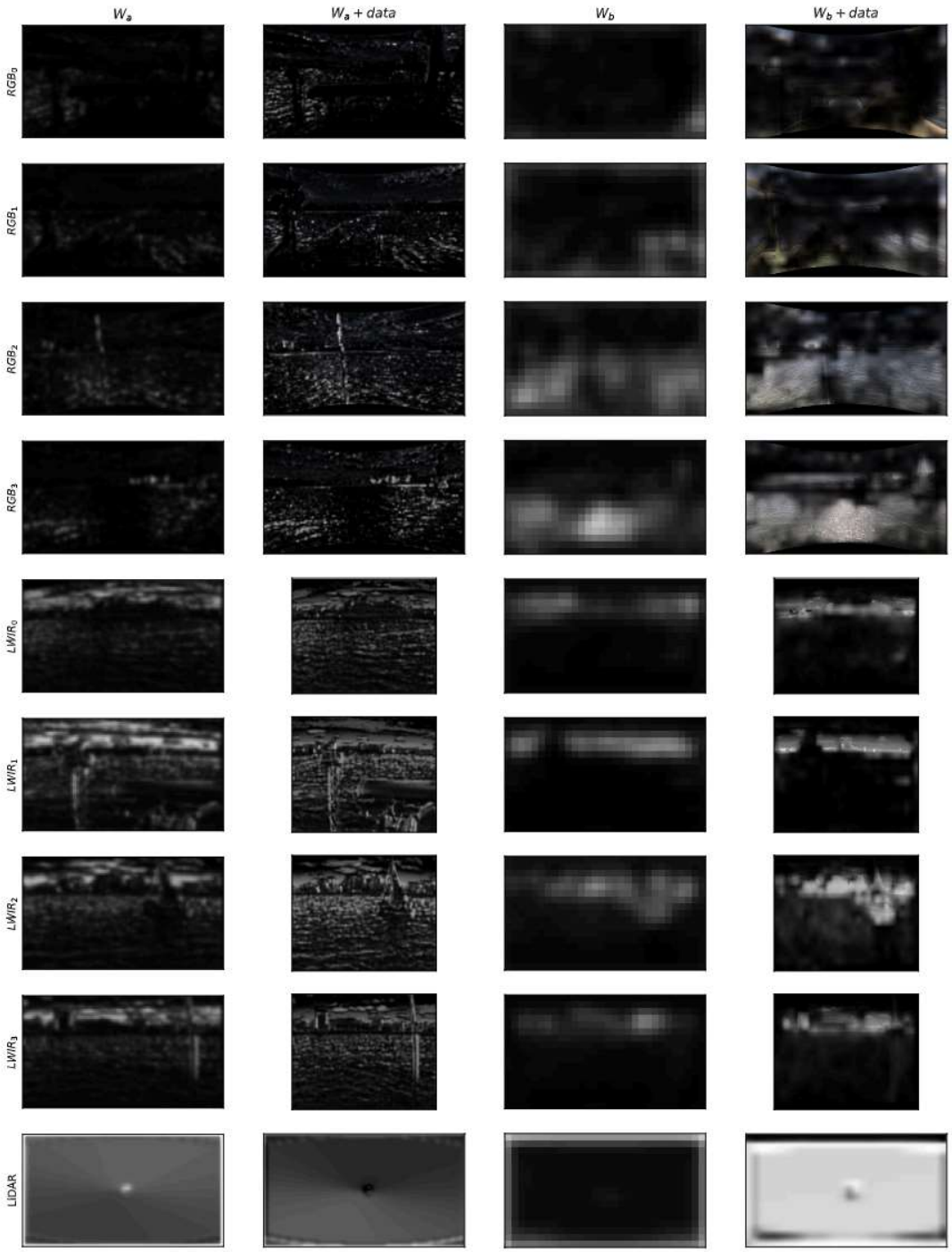

Fig. 14. Attention maps W_{probe} calculated from the two cross-attention modules, that operate on different input feature resolutions (fig. 11). The produced score map demonstrate the ability of the model to focus on salient regions that correspond to a BEV region (here to the regions highlighted by red color in fig. 5) discarding the non-relevant areas in feature fusion.