2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

ASY-VRNet: Waterway Panoptic Driving Perception Model based on Asymmetric Fair Fusion of Vision and 4D mmWave Radar

Runwei Guan[1,2,3,4 †], Shanliang Yao[1,2,3,4 †], Ka Lok Man[3], Xiaohui Zhu[3], Yong Yue[3], Jeremy Smith[1], Eng Gee Lim[3], *Senior Member, IEEE*, Yutao Yue[5,2,4 *]

***Abstract*— Panoptic Driving Perception (PDP) is critical for the autonomous navigation of Unmanned Surface Vehicles (USVs). A PDP model typically integrates multiple tasks, necessitating the simultaneous and robust execution of various perception tasks to facilitate downstream path planning. The fusion of visual and radar sensors is currently acknowledged as a robust and cost-effective approach. However, most existing research has primarily focused on fusing visual and radar features dedicated to object detection or utilizing a shared feature space for multiple tasks, neglecting the individual representation differences between various tasks. To address this gap, we propose a pair of Asymmetric Fair Fusion (AFF) modules with favorable explainability designed to efficiently interact with independent features from both visual and radar modalities, tailored to the specific requirements of object detection and semantic segmentation tasks. The AFF modules treat image and radar maps as irregular point sets and transform these features into a crossed-shared feature space for multitasking, ensuring equitable treatment of vision and radar point cloud features. Leveraging AFF modules, we propose a novel and efficient PDP model, ASY-VRNet, which processes image and radar features based on irregular super-pixel point sets. Additionally, we propose an effective multi-task learning method specifically designed for PDP models. Compared to other lightweight models, ASY-VRNet achieves state-of-the-art performance in object detection, semantic segmentation, and drivable-area segmentation on the WaterScenes benchmark. Our project is publicly available at https://github.com/GuanRunwei/ASY-VRNet.**

I. Introduction

With the rapid and exhilarating development of artificial intelligence and sophisticated perception sensors, USVs demonstrate immensely promising value in channel monitoring, water-quality assessment, water-surface rescue operations, water-surface transportation, and geological prospecting [1][2]. As one of the most crucial and foundational modules, perception is vital for the autonomous and efficient navigation of USVs. Currently, Panoptic Driving Perception (PDP) is regarded as an extraordinarily effective paradigm for comprehensive environmental perception, typically relying on advanced vision sensors. PDP operates based on multi-task robust perception for different driving areas, aiming at the simultaneous instance-level perception of objects and pixel-level recognition of drivable areas [3][4][5], which facilitates the comprehensive understanding of the environment. In contrast to integrating multiple single-task models, PDP models can significantly reduce memory usage and dramatically enhance inference speed. Moreover, through meticulously designed and well-coordinated multi-task training patterns, they effectively improve the performance of individual tasks, demonstrating both impressive efficiency and remarkable accuracy.

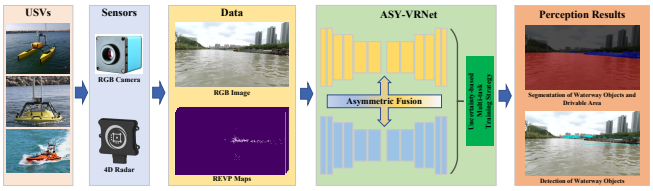

Fig. 1. The overview of our proposed methods. It contains five parts, USVs, sensors (monocular camera and 4D radar), data perceived by sensors, ASY-VRNet, multi-task training strategy and perception results.

However, as illustrated in Fig. 2, purely visual solutions often prove unreliable in various aquatic environments, such as (a) low-light conditions, (b) water droplets on the lens, (c) strong reflections on the water surface, (e) dense water fog, and (f) small objects. Currently, the fusion of visual information with 4D millimeter-wave radar (4D radar) is regarded as a promising, reliable, and cost-effective approach. As an all-weather perception sensor, 4D radar remains unaffected by adverse weather conditions and provides denser point cloud information compared to 3D radar, despite being susceptible to multi-path clutter (Fig. 2 (d)). Numerous studies have explored feature-level fusion of radar and vision for environmental perception, primarily aiming to enhance object detection performance [6][2][7][8]. Nevertheless, enabling region-level object detection and pixel-level semantic segmentation tasks through the effective utilization of vision and radar deep features to mutually enhance their performance remains a challenge. Specifically, radar point clouds are typically sparse and irregular, and visual objects often exhibit similar irregularities, suggesting that conventional convolution-based modules, which rely on 2D rectangular structures, may not effectively model both modalities. Additionally, region-

This work is partially supported by the XJTLU AI University Research Centre and Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU. Also, it is partially funded by the Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004) as well as funding: XJTLU-REF-21-01-002, XJTLU- RDF-22-01-062, and XJTLU Key Program Special Fund (KSF-A-17). This work received financial support from Jiangsu Industrial Technology Research Institute (JITRI) and Wuxi National Hi-Tech District (WND).

†Runwei Guan and Shanliang Yao contribute equally.
[1]Department of EEE, University of Liverpool, Liverpool, UK; [2]Institute of Deep Perception Technology, JITRI, Wuxi, China; [3]SAT, Xi'an Jiaotong-Liverpool University, Suzhou, China; [4]XJTLU-JITRI Academy of Industrial Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China; [5]Thrust of Artificial Intelligence and Thrust of Intelligent Transportation, HKUST (GZ), Guangzhou, China.
*Corresponding author: yutaoyue@hkust-gz.edu.cn

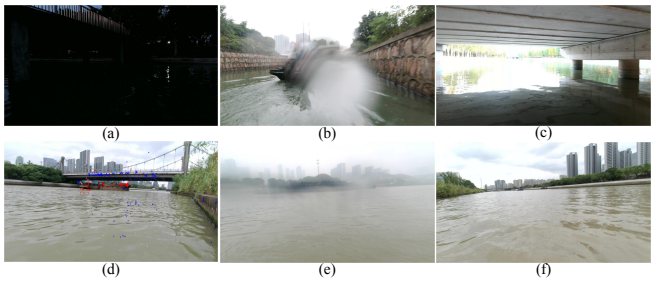

Fig. 2. Several challenging scenes in waterway perception: (a) dark environment, (b) camera malfunction, (c) strong light, (d) radar clutter, (e) adverse weather and (f) small objects.

level detection and pixel-level segmentation have different feature representation and optimization requirements, where the shared fusion modules [6][7][2] usually cannot help these two types of tasks achieve their respective optimal performances. Therefore, we propose a cephalocaudal feature structure that can impartially and consistently treat visual and radar point cloud features as sets of irregular pixel points during feature extraction, alignment, and fusion. Additionally, we design two specialized vision-radar fusion modules aimed at maximizing the performance of both detection and segmentation tasks.

Building on above, we concentrate on robust and high-performance panoptic driving perception for waterway autonomous driving and our contributions are as follows:

1) We propose a robust PDP model for waterway named ASY-VRNet. ASY-VRNet utilizes a full Contextual-Clustering (CoC) architecture [9], including both the backbone and neck, treating image and radar objects equitably as irregular point sets.
2) Drawing from prior and theoretical principles, we develop a pair of effective fusion methods for vision-radar integration, termed Asymmetric Fair Fusion (AFF). AFF takes into account the distinct characteristics of different perception tasks and designs the fusion processes based on the unique features of images and radar maps, ensuring an unbiased approach to both modalities. AFF enhances explainability and can be employed as a plug-and-play module to improve the performance of any vision-radar fusion network.
3) Inspired by homoscedastic uncertainty [10], we devise a multi-task training strategy that leverages the inherent uncertainty of perception tasks during the training of PDP models.

II. Related Works

A. Panoptic Driving Perception in Autonomous Driving

Panoptic driving perception (PDP) is crucial in autonomous driving systems for UGVs, USVs, and UAVs. PDP models are primarily responsible for obstacle detection and drivable area segmentation to facilitate downstream path planning. Current PDP models can be categorized into vision-based and fusion-based models. Vision-based models, such as YOLOP [3] and YOLOPv2 [4], both based on YOLO architecture, are capable of detecting traffic participants, identifying lanes, and segmenting drivable areas simultaneously. HybridNets [5], utilizing EfficientNet [11] as the backbone, combines this with an effective multi-task training strategy. In waterway perception, sensor-based models primarily involve the fusion of vision and radar. Mask-VRDet [2] employs a dual graph fusion (DGF) method to integrate image and radar features. Achelous [6] and Achelous++ [6] can perform five tasks concurrently, adopting both feature-level and proposal-level fusion strategies.

B. Multi-Task Learning Strategies in Computer Vision

Multi-task learning (MTL) is a fundamental and challenging problem in deep learning. Balancing the loss of various tasks during training based on their loss values and characteristics presents an intriguing proposition. For related tasks, designing appropriate MTL strategies can effectively enhance the representation of shared features, thereby improving performance across different tasks. GradNorm [12] is a classical MTL method that scales the loss of different tasks to a similar magnitude. Dynamic Weight Averaging (DWA) [13] balances the learning paces of various tasks. Sener *et al.* [14] approach MTL as a multiple objective optimization (MOO) problem, finding the Pareto optimal solution among various tasks. Kendall *et al.* [10] propose an uncertainty-based MTL method, estimating the weight of various tasks through Gaussian likelihood estimation.

III. ASY-VRNet

This section presents a detailed exposition of the design of ASY-VRNet, organized into six parts corresponding to the major modules of the model.

A. Alignment and Pre-processing of Perception Data

Given that all tasks are based on the camera plane, we initially project the 3D radar point cloud onto the camera plane through a coordinate system transformation.

After projecting the radar point clouds onto the camera plane, each radar point encapsulates features such as the object's range, elevation, velocity, and reflected power. Leveraging these characteristics, we design a radar data representation known as REVP maps, which is a 4-channel image-like feature map.

B. Vision-Radar-based Contextual Clustering (VRCoC)

Vision-**R**adar-based **Co**ntextual **C**lustering (VRCoC) constitutes the backbone of ASY-VRNet, facilitating hierarchical feature extraction across four stages. As illustrated in Fig. 3, VRCoC is a dual-branch backbone. Each branch consists of individual blocks formed by the combination and stacking of two fundamental modules: Point Reducer and Contextual Clustering. These branches are specifically designed to extract features from images and radar REVP maps, respectively.

Feature Preparation. Given an image $I \in \mathbb{R}^{3\times w\times h}$ and a REVP map $R \in \mathbb{R}^{4\times w\times h}$, we assign the coordinate of each pixel in the RGB image $I_{i,j}$ as $\left[\frac{i}{w}-0.5, \frac{j}{h}-0.5\right]$.

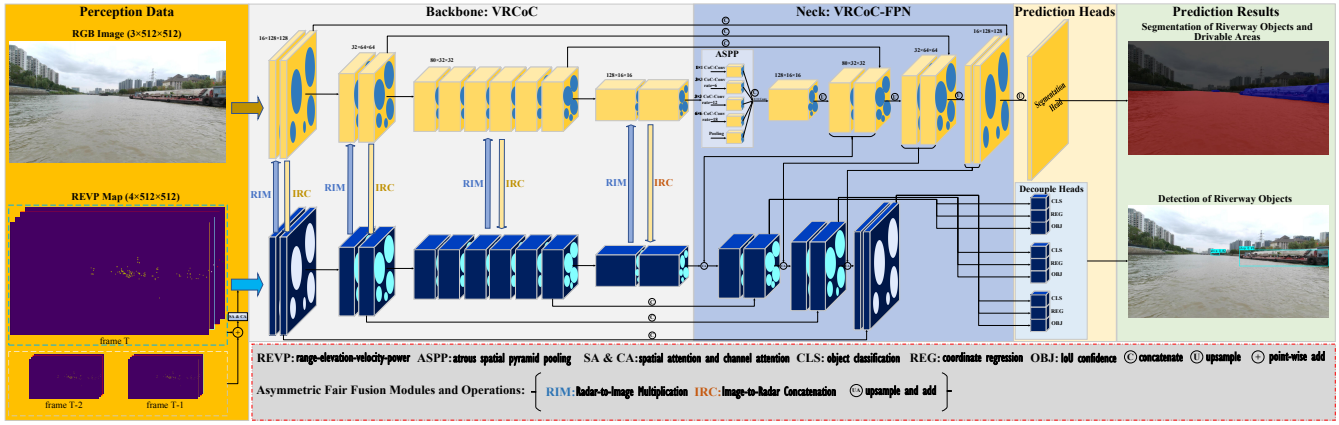

Fig. 3. The architecture of our proposed ASY-VRNet. It contains five parts, perception data, VRCoC, VRCoC-FPN, prediction heads and Asymmetric Fair Fusion modules (AFF), including RIM and IRC. Each stage of VRCoC has 2, 2, 6, 2 stacking blocks. VRCoC, AFF and VRCoC-FPN (Feature Pyramid Network) are three dedicated designed components in this paper.

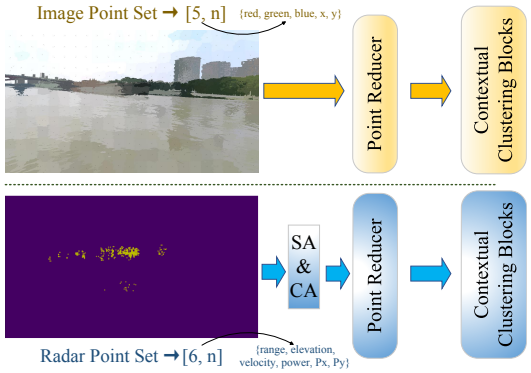

Fig. 4. The first stage of VRCoC, including image-like point sets (image and radar), point reducer and contextual clustering blocks.

Similarly, the coordinate of each element in the REVP map $R_{i,j}$ is assigned as $\left[\frac{i}{w}-0.5, \frac{j}{h}-0.5\right]$. Consequently, the RGB image is transformed into a set of points $IP \in \mathbb{R}^{5\times n}$ while the REVP map is transformed into a set of points $RP \in \mathbb{R}^{6\times n}$, where $n = w \times h$ is the point count per channel in both the RGB image and the REVP map. Each point in the RGB image encompasses color features (3 channels) and positional features (2 channels). For the REVP map, each point includes radar-captured object features (4 channels) and positional features (2 channels). Due to radar interference from multipath effects and surface clutter, numerous non-object clutter points may appear. To mitigate the impact of clutter features on model optimization, we incorporate dual attention mechanisms for both spatial and channel dimensions before feeding the REVP map into the Point Reducer module (Fig. 4). This allows the model to adaptively adjust its focus on different positions and feature channels of the REVP map. Our channel attention is based on the Efficient Channel Attention (ECA) module [15], while the spatial attention utilizes Deformable Convolution [16], effectively modeling the irregular features of point clouds.

Point Reducer. Based on the sets of points $IP \in \mathbb{R}^{5\times n}$ and $RP \in \mathbb{R}^{6\times n}$ obtained, we proceed with feature extraction. As depicted in Fig. 4, the first step involves the Point Reducer, which reduces the number of points. In this step, following the Contextual Clustering (CoC) methodology [9], we evenly select anchors in the feature space and concatenate the nearest k centers of points. Subsequently, a linear feed-forward module is employed to transform the feature map dimensions to d.

Contextual Clustering. Based on the feature points of the image $IP \in \mathbb{R}^{d\times n}$ and the REVP map $RP \in \mathbb{R}^{d\times n}$ at the same stage, we group feature points into several clusters based on the cosine similarity between the features of the points and the clustering centers. The clustering centers are evenly selected using the SLIC algorithm [17]. After assigning points to their respective centers, feature aggregation is applied upon the similarities between the clustering points and the clustering center. Assuming there are m clustering points in a cluster, the similarity matrix between the clustering points and the clustering center is denoted as $s \in \mathbb{R}^m$. We then map these feature points into an aggregation space, so $IP \in \mathbb{R}^{d\times m}$ and $RP \in \mathbb{R}^{d\times m}$ are transformed to $IP \in \mathbb{R}^{\hat{d}\times m}$ and $RP \in \mathbb{R}^{\hat{d}\times m}$, respectively, where $\hat{d}$ is the dimension of the feature points in the aggregation space. Within each cluster in the aggregation space, a clustering center v_c is similarly proposed based on the SLIC algorithm. Therefore, the aggregated feature of points $f \in \mathbb{R}^{\hat{d}\times m}$ is presented in Equation 1.

$$
\begin{aligned}
f &= \frac{1}{C}\Big(v_c + \sum_{i=1}^{m} \sigma(\alpha s_i + \beta) * v_i\Big), \\
s.t., C &= 1 + \sum_{i=1}^{m} \sigma(\alpha s_i + \beta),
\end{aligned}
\tag{1}
$$

where α and β are learnable parameters representing the scale and shift ratio of the similarity. σ is the sigmoid function to scale the similarity to $(0, 1)$. v_i denotes the ith points in the aggregation space. C is a normalization factor.

Then the aggregated feature f is dispatched to each feature point in the cluster according to the similarity. For each feature point p_i, the dispatch step for updating is presented in Equation 2,

$$\hat{p_i} = p_i + FF\Big(\sigma(\alpha s_i + \beta) * f\Big), \quad \hat{p_i} \in \mathbb{R}^{d \times n}, \tag{2}$$

where p_i represents the i^{th} feature point and $\hat{p_i}$ represents the updated i^{th} feature point. σ is the sigmoid function. FF is the feed-forward module based on fully-connected layers, which transforms the dimension $\hat{d}$ back to d.

Based on the above, we obtain the feature maps of the image $\hat{IP}$ and radar $\hat{RP}$ updated by the point reducer and contextual clustering in CoC. Inspired by multi-head self-attention, we divide the channel of $\hat{IP}$ and $\hat{RP}$ into h parts, and each part denotes one head. Each head $head_i$ is weighted individually and concatenated to other heads. After that, we concatenate all heads along the dimension of channels. Multi-head operation can make the network adaptively attach importance to features. The process of the multi-head operation is shown in Equation 3.

$$\begin{aligned} \hat{IP}^{'} &= [head_1^{IP} W_1^{IP}, head_2^{IP} W_2^{IP}, \ldots, head_h^{IP} W_h^{IP}], \\ \hat{RP}^{'} &= [head_1^{RP} W_1^{RP}, head_2^{RP} W_2^{RP}, \ldots, head_h^{RP} W_h^{RP}]. \end{aligned} \tag{3}$$

C. Asymmetric Fair Fusion Modules

The Asymmetric Fair Fusion (AFF) modules are a set of bidirectional fusion mechanisms designed for the mutual integration and enhancement of visual and radar features. They are structured asymmetrically and consist of two primary components: Image-Radar Concatenation (IRC) and Radar-Image Multiplication (RIM).

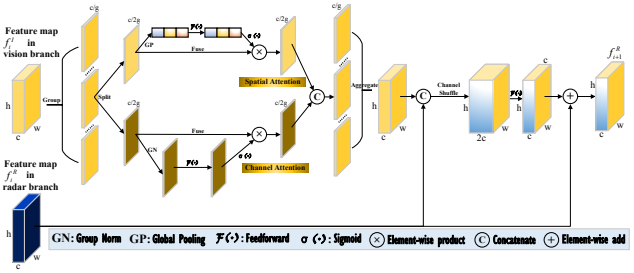

Fig. 5. The structure of Image-Radar Concatenation (IRC).

Image-Radar Concatenation (IRC) is designed to enhance object detection representation. It is widely recognized that image-based object detection results are not directly influenced by pixel brightness or color. For instance, a luminous area does not necessarily indicate the location of an object. Neural network models typically learn from a vast number of images to identify various feature combinations for object localization. However, 4D radar significantly improves this process. 4D radar can capture denser point clouds of objects than ordinary radar, regardless of whether the object is moving or stationary. This capability allows the point cloud of 4D radar to help the neural network model anchor the general area of the object early in training, thereby accelerating the convergence of object detection. In adverse weather and low-light environments, the 4D radar point cloud can compensate for the lack of visual features, reducing the likelihood of missed detections.

Based on the aforementioned principles, we propose the Image-to-Radar Concatenation (IRC) module. As depicted in Fig. 5, let's consider a feature map $f_i^I \in \mathbb{R}^{c \times h \times w}$ in the vision branch and a feature map $f_i^R \in \mathbb{R}^{c \times h \times w}$ in the radar branch. f_i^I is initially divided into g segments, where each segment is denoted as $f_{i_j}^I \in \mathbb{R}^{\frac{c}{g} \times h \times w}$. These segments undergo further processing: one branch focuses on spatial attention while the other on channel attention, following the principles of shuffle attention [18].

For channel attention, given the input feature map $f_{i_{j-c}}^I \in \mathbb{R}^{\frac{c}{2g} \times h \times w}$, as expressed in Equation 4, $f_{i_{j-c}}^I$ undergoes global average pooling to capture its global representation. Subsequently, non-linear and sigmoid functions are applied to assess the importance of each channel. Finally, the channel importance weight is applied to $f_{i_{j-c}}^I$, resulting in the feature map with channel attention $\hat{f}_{i_{j-c}}^I \in \mathbb{R}^{\frac{c}{2g} \times h \times w}$.

$$\begin{aligned} f_{i_{j-c-1}}^I &= \sigma\Big(W_c \cdot GP(f_{i_{j-c}}^I) + b_c\Big), \quad f_{i_{j-c-1}}^I \in \mathbb{R}^{\frac{c}{2g} \times 1 \times 1}, \\ \hat{f}_{i_{j-c}}^I &= f_{i_{j-c-1}}^I * f_{i_{j-c}}^I, \quad \hat{f}_{i_{j-c}}^I \in \mathbb{R}^{\frac{c}{2g} \times h \times w}, \end{aligned} \tag{4}$$

where GP denotes global average pooling. W_c is the learnable weight in the non-linear feed-forward module while b_c is the learnable bias. σ is the sigmoid function. $*$ denotes element-wise multiplication.

For the spatial attention, given the input feature map $f_{i_{j-s}}^I \in \mathbb{R}^{\frac{c}{2g} \times h \times w}$, as Equation 5 presents, $f_{i_{j-s}}^I$ is first normalized by group, then processed by a non-linear feed-forward module and a sigmoid function to measure the spatial importance. Finally, the spatial importance is multiplied with the input feature map $f_{i_{j-s}}^I$ to obtain the feature map with spatial attention $\hat{f}_{i_{j-s}}^I \in \mathbb{R}^{\frac{c}{2g} \times h \times w}$.

$$\begin{aligned} f_{i_{j-s-1}}^I &= \sigma\Big(W_s \cdot GN(f_{i_{j-s}}^I) + b_s\Big), \quad f_{i_{j-s-1}}^I \in \mathbb{R}^{\frac{c}{2g} \times h \times w}, \\ \hat{f}_{i_{j-s}}^I &= f_{i_{j-s-1}}^I * f_{i_{j-s}}^I, \quad \hat{f}_{i_{j-s}}^I \in \mathbb{R}^{\frac{c}{2g} \times h \times w}, \end{aligned} \tag{5}$$

where GN denotes group-norm. W_s is the learnable weight in the non-linear feed-forward module while b_s is the learnable bias. σ is the sigmoid function while $*$ denotes element-wise multiplication.

After that, we concatenate $\hat{f}_{i_{j-s}}^I$ and $\hat{f}_{i_{j-c}}^I$ and get the combination of the feature map $\hat{f}_{i_{j-sc}}^I$ with both channel and spatial attention. Aggregation (concatenation) of g $\hat{f}_{i_{j-sc}}^I$ is implemented to get the initial feature map with channel and spatial attention $f_{i_{sc}}^I \in \mathbb{R}^{c \times h \times w}$. We concatenate $f_{i_{sc}}^I$ and $f_i^R \in \mathbb{R}^{c \times h \times w}$ along the channel dimension (Equation 6).

$$f_i^{IR} = [f_{i_{sc}}^I, f_i^R], \quad f_i^{IR} \in \mathbb{R}^{2c \times h \times w}, \tag{6}$$

where $[\cdot]$ is the concatenation operation.

After that, the channel shuffle is exerted to enhance the interaction among features, followed by a feed-forward module, reducing the channel dimension. Finally, a long residual path is added and we get the feature map f_{i+1}^R updated by *IRC* in the radar branch. The whole process is shown in Equation 7.

$$
\begin{aligned}
f_i^{IR} &= S(f_i^{IR}), \quad f_i^{IR} \in \mathbb{R}^{2c\times h\times w}, \\
\hat{f}_i^{IR} &= W_f \cdot f_i^{IR} + b_f, \quad \hat{f}_i^{IR} \in \mathbb{R}^{c\times h\times w}, \\
f_{i+1}^{R} &= \hat{f}_i^{IR} + f_i^{R}, \quad f_{i+1}^{R} \in \mathbb{R}^{c\times h\times w},
\end{aligned} \tag{7}
$$

where $S(\cdot)$ denotes the channel shuffle. W_f is the learnable weight in the non-linear feed-forward module while b_f is the learnable bias.

Radar-Image Multiplication (RIM) is to enhance the representation for image segmentation as radar point clouds can be seen as sparse masks of objects. RIM is based on the formula of brightness and contrast adjustment (Equation 8).

$$
g(i,j) = \alpha f(i,j) + \beta, \tag{8}
$$

where $f(i,j)$ is the pixel in the original image while $g(i,j)$ is the pixel after adjustment. α is the gain to adjust the image contrast while β is the bias to control the image brightness. Based on the above, we intend to use features in the radar branch to focus on and enhance the features in the vision branch at same positions.

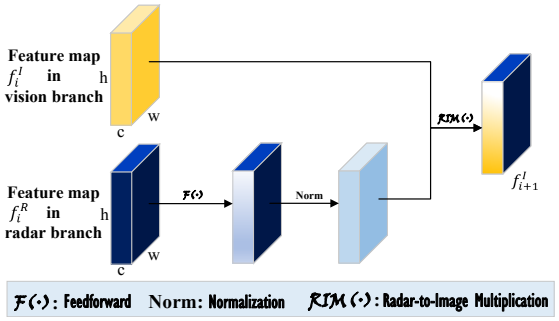

Fig. 6. The architecture of Radar-to-Image Multiplication (RIM).

Fig. 6 presents the architecture of the Radar-Image Multiplication (RIM) module. Assuming a feature map $f_i^I \in \mathbb{R}^{c\times h\times w}$ in the vision branch and a feature map $f_i^R \in \mathbb{R}^{c\times h\times w}$ in the radar branch, as shown in Equation 9, f_i^R first undergoes a feed-forward module and normalization to produce the feature map $\hat{f}_i^R$. Following Equation 8, the parameters $\alpha = 1 + \hat{f}_i^R$ and $\beta = \gamma * \hat{f}_i^R$ are utilized to enhance the corresponding positions in the vision branch. This process results in the image feature map f_{i+1}^I, which incorporates cross attention from the radar feature map.

$$
\begin{aligned}
\hat{f}_i^R &= Norm(W_r \cdot f_i^R + b_r), \quad \hat{f}_i^R \in \mathbb{R}^{c\times h\times w}, \\
f_{i+1}^I &= (1 + \hat{f}_i^R) \cdot f_i^I + \gamma * \hat{f}_i^R, \quad f_{i+1}^I \in \mathbb{R}^{c\times h\times w},
\end{aligned} \tag{9}
$$

where $Norm$ is the normalization operation. W_r is the learnable weight and b_r is the learnable bias in the feed-forward module. γ is a learnable coefficient.

D. Dual Feature Pyramid Networks

To maintain consistency in the FPN structure with the backbone, we continue to employ Contextual Clustering (CoC) as the fundamental unit across all stages of FPN, termed VRCoC-FPN. Illustrated in Fig. 3, VRCoC-FPN retains the dual-branch architecture akin to VRCoC. In the vision branch, each stage incorporates an Atrous Spatial Pyramid Pooling (ASPP) module [19] to enhance the receptive field across multiple scales. Skip connections are employed within VRCoC-FPN to facilitate multi-scale feature fusion. Meanwhile, in the radar branch, each stage integrates feature maps from the corresponding stage in the vision branch to enhance resolution for object detection.

E. Predictions Heads

The model comprises two distinct prediction heads: one for semantic segmentation and another for object detection. The segmentation head consists of $C_{seg}+1$ channels, where C_{seg} denotes the number of segmentation categories, and 1 represents the background. For object detection, we utilize decoupled heads, inspired by YOLOX [20], to independently predict bounding box coordinates, object category, and confidence score. Besides, ASY-VRNet is anchor-free and employs SimOTA [20] for positive sample matching.

F. Multi-task Optimization Strategy

Given the significant disparity in loss magnitudes between object detection and semantic segmentation, we adopt a multi-task loss approach inspired by Kendall *et al.* [10], which is based on homoscedastic uncertainty. Homoscedastic uncertainty, a subset of aleatoric uncertainty, pertains to inherent data randomness and unexplainable information. ASY-VRNet addresses two primary tasks: object detection and semantic segmentation. Object detection includes the loss of bounding box coordinates, confidence and object category, which are one regression and two classification tasks. Semantic segmentation is a pixel-level classification task. Therefore, we can consider these two tasks as a combination of regression and classification sub-tasks. We define the loss of our ASY-VRNet model as $L(W,\sigma_1,\sigma_2,\sigma_3,\sigma_4)$, which can be written as,

$$
\begin{aligned}
&L(W,\sigma_1,\sigma_2,\sigma_3,\sigma_4) \\
&= \sum_{i=1}^{3} \frac{1}{\sigma_i^2} L_i(W) + \frac{1}{\sigma_4^2} L_4(W) + \sum_{k=1}^{4} log\sigma_k,
\end{aligned} \tag{10}
$$

where σ_1, σ_2, σ_3, σ_4 respectively represent the uncertainty of the data for 4 sub-tasks in panoptic perception: object classification, object confidence score, pixel classification and bound box regression. $log\sigma_k$ is the regularization term. If σ_k became larger, the weight of $L_k(W)$ would be smaller.

IV. Experiments

A. Experimental Settings

We train and evaluate ASY-VRNet on the WaterScenes dataset [21], including 54,120 frames and seven categories in various waterway scenarios. We train all models in experiments for 100 epochs with a batch size of 16. We adopt Stochastic Gradient Descent with Momentum (SGDM) as the optimizer. The weight decay is 5e-4 while the momentum is 0.937. We adopt a cosine learning rate scheduler with an initial learning rate of 1e-2. Both images and REVP maps are resized as 320×320 (px) during the training. Furthermore,

TABLE I. Comparison with Other Models on Object Detection

Models	Modalities	Params (M)	FLOPs (G)	$mAP_{50\text{-}95}$	$AR_{50\text{-}95}$
Single-Task Models					
YOLOv4-T [20]	$\mathbb{V}$	5.89	4.04	13.1	20.2
YOLOv7-T [22]	$\mathbb{V}$	6.03	33.3	37.3	43.7
YOLOX-T [20]	$\mathbb{V}$	5.04	3.79	39.4	43.0
YOLOv8-N [23]	$\mathbb{V}$	3.01	2.05	41.9	44.0
CRFNet [24]	$\mathbb{V}$+$\mathbb{R}$	23.54	-	41.8	44.5
Multi-Task Models					
YOLOP [3]	$\mathbb{V}$	7.90	18.60	37.9	43.5
HybridNets [5]	$\mathbb{V}$	12.83	15.60	39.1	44.2
Achelous [6]	$\mathbb{V}$+$\mathbb{R}$	3.49	3.04	41.5	45.6
ASY-VRNet	$\mathbb{V}$+$\mathbb{R}$	**4.12**	**3.26**	**42.8**	**46.3**

TABLE II. Comparison with Other Models on Semantic Segmentation (Object and Drivable-area)

Models	Modalities	Params (M)	FLOPs (G)	$mIoU_o^1$	$mIoU_d^2$
Single-Task Models					
Segformer-B0 [25]	$\mathbb{V}$	3.71	5.29	73.5	99.4
DeepLabV3+ [26]	$\mathbb{V}$	5.81	20.60	71.6	99.2
PSPNet [27]	$\mathbb{V}$	2.38	2.30	69.4	99.0
Multi-Task Models					
YOLOP [3]	$\mathbb{V}$	7.90	18.60	-	99.0
HybridNets [5]	$\mathbb{V}$	12.83	15.60	-	98.8
Achelous [6]	$\mathbb{V}$+$\mathbb{R}$	3.49	3.04	70.6	99.5
ASY-VRNet	$\mathbb{V}$+$\mathbb{R}$	**4.12**	**3.26**	**74.7**	**99.6**

1. mIoU of objects; 2. mIoU of drivable area.

we adopt Exponential Moving Average (EMA) and Mixed Precision (MP). For the test, we choose mAP, AP and AR as metrics to evaluate the detection performances, while mIoU is the semantic segmentation metric. All training and test works are implemented on one TITAN RTX GPU.

B. Comparison on Object Detection

As TABLE I shows, for object detection, we compare performances of single-task models, multi-task models, vision-based models, and radar-vision fusion models, which mainly include models with similar orders of magnitude of parameters. Generally speaking, our ASY-VRNet achieves state-of-the-art performances whatever $mAP_{50\text{-}95}$ and $AR_{50\text{-}95}$ with generally fewer parameters and FLOPs. Exactly, compared with another fusion-based PDP multi-task model Achelous (MV-GDP-X-PN), our ASY-VRNet exceeds about 1.3 $mAP_{50\text{-}95}$ while 0.7 $AR_{50\text{-}95}$. For another two vision-based multi-task models with more parameters and FLOPs, YOLOP and HybridNets, our ASY-VRNet achieves 3.7 and 4.9 higher $mAP_{50\text{-}95}$. Furthermore, our ASY-VRNet gets the best performances when compared with single-task models. Notably, ASY-VRNet outperforms CRFNet 1 $mAP_{50\text{-}95}$ with about 19 million fewer parameters. From another perspective, we find that fusion-based models generally obtain better detection recall than vision-based models, which means a lower miss-detection rate based on fusion-based perception methods. Further, as TABLE VI shows, ASY-VRNet outperforms Achelous under several adverse situations.

C. Comparison on Semantic Segmentation

TABLE II presents the semantic segmentation results of various models on the segmentation of objects and drivable areas. It is apparent that our ASY-VRNet achieves the best performance among all single-task models and multi-task models. Specifically, ASY-VRNet outperforms Achelous (MV-GDP-X-PN) by about 4.1 mIoU when segmenting objects. Although YOLOP and HybridNets have more parameters, their performances are still worse than ASY-VRNet, which proves the effectiveness of our model architecture. Besides, ASY-VRNet achieves 1.2 higher mIoU than Segformer-B0 with fewer FLOPs.

D. Ablation Experiments

To demonstrate the efficacy of our ASY-VRNet's modules, we conduct ablation experiments. As summarized in TABLE III, various modules including RIM, IRC, the fusion branch in the neck (neck fusion), CoC-FPN, decouple detection head, and multi-frame radar data are evaluated. Notably, for object detection, IRC has the most significant impact, leading to a decrease in mAP of approximately 1.2%. Additionally, replacing CoC-FPN with a conventional convolutional FPN resulted in a 1.0% drop in mAP. Furthermore, the neck fusion branch, decouple detection head, and multi-frame data each contributed to enhancing object detection performance to varying degrees. For semantic segmentation, RIM is found to be the most critical component. Notably, replacing CoC-FPN with a conventional FPN causes a 1.8% decline in mIoU. CoC-FPN also demonstrates some improvement in semantic segmentation. Based on these findings, we observe that removing CoC-FPN hurt both object detection and semantic segmentation. This suggests that the consistency between the feature structure of the FPN and the backbone is essential.

E. Comparison on Multi-Task Training

As shown in TABLE IV, we utilize four multi-task training techniques to train our ASY-VRNet. The techniques include task-joint methods containing uncertainty weighting, manual weighting, GradNorm [12] and MGDA [14]. Notably, the uncertainty-based training approach delivers exceptional overall performance, with the lowest detection loss and optimized segmentation results. When training object detection independently, the mAP metric is unsatisfactory. It's noteworthy that manually tuning the weights of the four sub-tasks is a challenging task, as its optimization performance falls short of the uncertainty-based training method. Furthermore, GradNorm and MGDA achieve competitive performances, but are still worse than our uncertainty-based training approach. In summary, our multi-task training strategy can effectively improve the performance of each individual task.

F. Experiments on Fusion Methods

We compare our proposed Asymmetric Fair Fusion (AFF) modules with several well-known fusion methods, including Multi-Head Cross Attention (MHCA), TokenFusion, and the fast fusion module in Achelous [6]. As shown in TABLE V, our AFF modules achieve state-of-the-art performance across

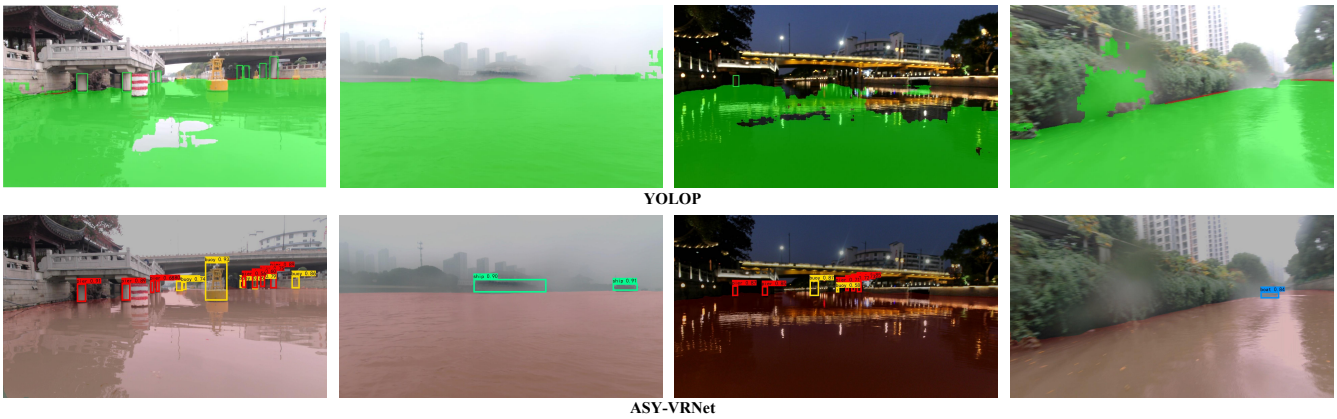

Fig. 7. Visualization of panoptic driving perception results predicted by YOLOP and ASY-VRNet, including scenarios of dense objects, dense fog, low light and droplets on the lens. Besides, the drivable area predicted by YOLOP is green while ASY-VRNet's is red.

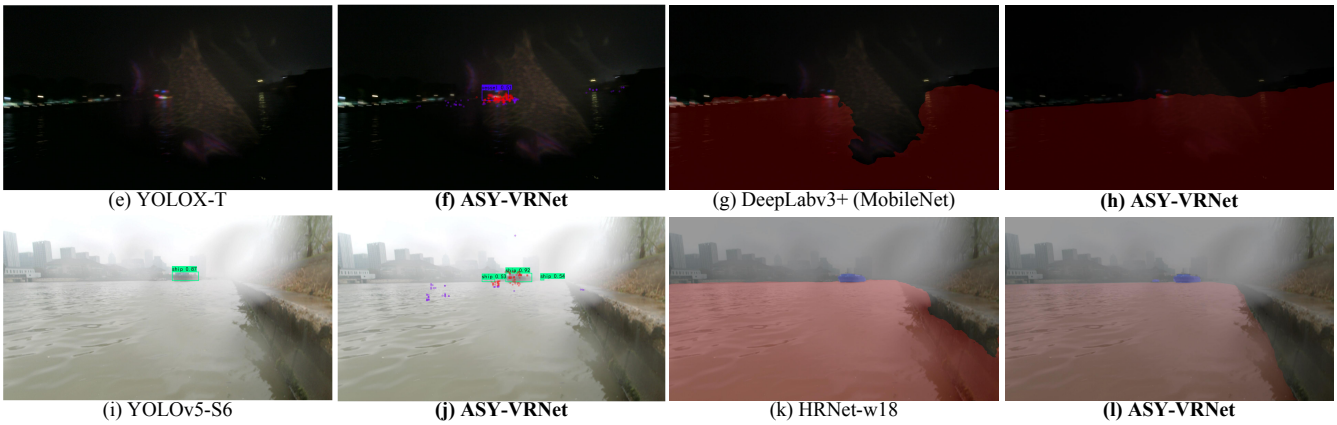

Fig. 8. Visualization of single-task models and ASY-VRNet, presenting distant vessel at night and interfered camera, and multiple ships on a foggy day. The first column is results of pure vision models while the second column presents the detection of our ASY-VRNet with 4D radar point clouds. The third column presents the semantic segmentation of pure vision models while the fourth column shows the segmentation of ASY-VRNet.

TABLE III. Ablation Experiments of ASY-VRNet

Methods	**mAP$_{50\text{-}95}$**	**mIoU$_o$**	**mIoU$_d$**
ASY-VRNet	**42.8**	**74.7**	**99.6**
-RIM	-	72.9 (↓ 1.8)	-
-IRC	41.6 (↓ 1.2)	-	
-neck fusion	42.4 (↓ 0.4)	-	
-decouple detection head	42.6 (↓ 0.2)	-	
-SA&CA on radar maps	42.6 (↓ 0.2)	74.4 (↓ 0.3)	99.5 (↓ 0.1)
multi-frame→single-frame	42.3 (↓ 0.5)	74.1 (↓ 0.6)	-
CoC-FPN → Conv-FPN	41.8 (↓ 1.0)	73.9 (↓ 0.8)	99.1 (↓ 0.5)

three perception tasks, surpassing other fusion methods. The performance of MHCA, which relies on the global receptive field, is not satisfactory. Similarly, the modal-agnostic fusion method TokenFusion shows a significant performance gap compared to our AFF modules, underscoring the importance of dedicated fusion methods tailored for various tasks.

G. Visualization and Analysis

We visualize the prediction results of our ASY-VRNet and other models as Fig. 7 and Fig. 8 present. We first select four representative samples under various scenarios, including dense and small objects, dense fog, low light and droplets on the lens. For the scenario containing dense small objects, we can find that our ASY-VRNet can nicely

TABLE IV. Results of Multi-Task Training Methods

Methods	Weights det_{bbox}	det_{conf}	det_{cls}	seg_{cls}	loss_{det}	loss_{seg}	**mAP$_{50\text{-}95}$**	**mIoU$_o$**	**mIoU$_d$**
Uncertainty Weighting	✓	✓	✓	✓	**3.273**	**0.301**	**42.8**	**74.7**	**99.6**
Manual	0.6	0.2	0.2	0.0	3.917	-	42.6	-	
Manual	0.5	0.2	0.2	0.1	3.578	0.355	42.3	74.4	99.6
Manual	0.25	0.25	0.25	0.25	4.231	0.322	42.0	73.9	99.3
Manual	0.1	0.2	0.2	0.5	5.241	0.352	41.6	74.6	99.6
Manual	0.0	0.0	0.0	1.0	-	0.303	-	74.6	99.6
GradNorm [12]	-	-	-	-	3.327	0.349	42.5	74.4	99.6
MGDA [14]	-	-	-	-	3.319	0.356	42.3	74.3	99.6

TABLE V. Comparison of Fusion Methods for Vision and Radar

Methods	**mAP$_{50\text{-}95}$**	**mIoU$_o$**	**mIoU$_d$**
AFF (ours)	**42.8**	**74.7**	**99.6**
MHCA [28]	40.4	68.2	97.5
TokenFusion [29]	41.2	68.0	97.8
Achelous [6]	42.1	72.0	99.5

TABLE VI. Comparison of Models under Adverse Situations

Models	**mAP^{da}**	**mIoU$_d^{da}$**	**mAP^{di}**	**mIoU$_d^{di}$**	**mAP^{sm}**	**mIoU$_o^{sm}$**
ASY-VRNet	**38.8**	**93.7**	**39.5**	**95.6**	**36.7**	**68.8**
Achelous	37.2	90.9	38.8	95.2	33.0	63.7

da: dark, **di**: dim, **sm**: small, **d**: drivable-area, **o**: object

detect all objects with high confidence scores while YOLOP misses considerable objects. For the objects behind the dense fog, YOLOP, unfortunately, misses both two ships while our ASY-VRNet successfully detects two moving ships. Moreover, the result of drivable-area segmentation by ASY-VRNet is better than YOLOP. For the scenario of low light, the drivable area predicted by YOLOP contains a lot of false-negative zones while our ASY-VRNet can better recognize them. For the fourth sample, we find that YOLOP can not perceive the object when some water droplets on the camera occluded the object, and it predicts many false positive drivable areas. In contrast, our ASY-VRNet can still capture the driving boat and smoothly segment the driving area. From another perspective, when compared with single-task models in Fig. 8, we can also observe the outstanding performances of ASY-VRNet in different scenarios.

V. Conclusions

In this paper, we propose ASY-VRNet, a PDP model for waterways that concurrently performs two distinct tasks. Our Asymmetric Fair Fusion (AFF) module efficiently integrates complementary features from each modality, enhancing the performance of both tasks simultaneously. Furthermore, we adopt a homoscedastic-uncertainty-based multi-task training method tailored for panoptic perception tasks, demonstrating its efficacy. ASY-VRNet treats both images and radar point clouds as irregular point sets, achieving competitive performance compared to other state-of-the-art single-task models. Moreover, it surpasses existing vision-based and fusion-based PDP models in overall performance.

References

[1] S. Yao, R. Guan, X. Huang, Z. Li, X. Sha, Y. Yue, E. G. Lim, H. Seo, K. L. Man, X. Zhu, and Y. Yue, "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review," IEEE Transactions on Intelligent Vehicles, pp. 1–40, 2023.

[2] R. Guan, S. Yao, L. Liu, X. Zhu, K. L. Man, Y. Yue, J. Smith, E. G. Lim, and Y. Yue, "Mask-vrdet: A robust riverway panoptic perception model based on dual graph fusion of vision and 4d mmwave radar," Robotics and Autonomous Systems, vol. 171, p. 104572, 2024.

[3] D. Wu, M.-W. Liao, W.-T. Zhang, X.-G. Wang, X. Bai, W.-Q. Cheng, and W.-Y. Liu, "Yolop: You only look once for panoptic driving perception," Machine Intelligence Research, vol. 19, no. 6, pp. 550–562, 2022.

[4] C. Han, Q. Zhao, S. Zhang, Y. Chen, Z. Zhang, and J. Yuan, "Yolopv2: Better, faster, stronger for panoptic driving perception," arXiv preprint arXiv:2208.11434, 2022.

[5] D. Vu, B. Ngo, and H. Phan, "Hybridnets: End-to-end perception network," arXiv preprint arXiv:2203.09035, 2022.

[6] R. Guan, S. Yao, X. Zhu, K. L. Man, E. G. Lim, J. Smith, Y. Yue, and Y. Yue, "Achelous: A fast unified water-surface panoptic perception framework based on fusion of monocular camera and 4d mmwave radar," in 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), pp. 182–188, IEEE, 2023.

[7] R. Guan, H. Zhao, S. Yao, K. L. Man, X. Zhu, L. Yu, Y. Yue, J. Smith, E. G. Lim, W. Ding, et al., "Achelous++: Power-oriented water-surface panoptic perception framework on edge devices based on vision-radar fusion and pruning of heterogeneous modalities," arXiv preprint arXiv:2312.08851, 2023.

[8] Y. Cheng, H. Xu, and Y. Liu, "Robust small object detection on the water surface through fusion of camera and millimeter wave radar," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15263–15272, 2021.

[9] X. Ma, Y. Zhou, H. Wang, C. Qin, B. Sun, C. Liu, and Y. Fu, "Image as set of points," in International Conference on Learning Representations, 2023.

[10] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7482–7491, 2018.

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

[12] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in International conference on machine learning, pp. 794–803, PMLR, 2018.

[13] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1871–1880, 2019.

[14] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," Advances in neural information processing systems, vol. 31, 2018.

[15] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11534–11542, 2020.

[16] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9308–9316, 2019.

[17] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 11, pp. 2274–2282, 2012.

[18] Q.-L. Zhang and Y.-B. Yang, "Sa-net: Shuffle attention for deep convolutional neural networks," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2235–2239, IEEE, 2021.

[19] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European conference on computer vision (ECCV), pp. 801–818, 2018.

[20] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," arXiv preprint arXiv:2107.08430, 2021.

[21] S. Yao, R. Guan, Z. Wu, Y. Ni, Z. Huang, R. W. Liu, Y. Yue, W. Ding, E. G. Lim, H. Seo, et al., "Waterscenes: A multi-task 4d radar-camera fusion dataset and benchmarks for autonomous driving on water surfaces," IEEE Transactions on Intelligent Transportation Systems, 2024.

[22] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7464–7475, 2023.

[23] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023.

[24] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in 2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF), pp. 1–7, IEEE, 2019.

[25] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," Advances in Neural Information Processing Systems, vol. 34, pp. 12077–12090, 2021.

[26] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, p. 833–851. Jan 2018.

[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890, 2017.

[28] D. Wu, W. Han, T. Wang, X. Dong, X. Zhang, and J. Shen, "Referring multi-object tracking," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14633–14642, 2023.

[29] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12186–12195, 2022.