Journal of Intelligent & Robotic Systems (2021) 102: 21
https://doi.org/10.1007/s10846-021-01395-1

REGULAR PAPER

Estimating Obstacle Maps for USVs Based on a Multistage Feature Aggregation and Semantic Feature Separation Network

Jingyi Liu[1] · Hengyu Li[1] · Jun Luo[1] · Shaorong Xie[1] · Yu Sun[2]

Received: 12 November 2020 / Accepted: 8 April 2021 / Published online: 24 April 2021

Abstract

Obstacle map estimation based on efficient semantic segmentation networks is promising for improving the environmental awareness of unmanned surface vehicles (USVs). However, existing networks perform poorly in challenging scenes with small obstacles, scenery reflections, boat wakes, and visual ambiguities caused by unfavorable weather conditions. In this paper, we address the small obstacle segmentation problem by learning representations of obstacles at multiple scales. An efficient multistage feature aggregation (MFA) module is proposed, which utilizes fully separable convolutions of different sizes to capture and fuse multiscale context information from different stages of a backbone network. In addition, a novel feature separation (FS) loss function based on Gaussian mixture model is presented, which encourages the MFA module to enforce separation among different semantic features, thereby providing a robust and discriminative representation in various challenging scenes. Building upon the MFA module and the FS loss function, we present a fast multistage feature aggregation and semantic feature separation network (FASNet) for obstacle map estimation of USVs. An extensive evaluation was conducted on a challenging public dataset (MaSTr1325). We validated that various lightweight semantic segmentation models achieved consistent performance improvement when our MFA module and FS loss function were adopted. The evaluation results showed that the proposed FASNet outperformed state-of-the-art lightweight models and achieved 96.71% mIoU and $>$ 1.5% higher obstacle-class IoU than the second-best network, while running over 58 fps.

Keywords Obstacle map estimation · Unmanned surface vehicles (USVs) · Semantic segmentation · Convolutional neural network (CNN)

1 Introduction

Recently, unmanned surface vehicles (USVs) have attracted a considerable amount of attention due to their portability, excellent concealment, and low cost [1–6]. These advantages make USVs ideal for oceanographic research, reconnaissance and patrolling missions in coastal waters and narrow marinas. However, in autonomous navigation, various obstacles may appear on the preplanned path of a USV, posing threats to the USV. Therefore, it is necessary for USVs to have an onboard unit for obstacle map estimation, adjust planned paths, and avoid collisions. Various sensors have been explored to handle this task, including sonar [7], radar [8], LIDAR [9], and cameras [1–5]. Among them, camera-based methods have become popular because cameras are lightweight and affordable and require minimal power consumption. These characteristics render cameras more suitable for USVs with limited payload and power consumption.

Reliable and efficient obstacle map estimation is a fundamental prerequisite for USVs to achieve autonomous navigation. Over the last decade, a variety of methods based on computer vision techniques have been proposed for maritime environment perception. Conventional methods [10–14] mostly utilize the horizon as guidance to reduce the search space of potential obstacles. These methods first detect a straight horizon line and then perform obstacle detection in the region below the horizon using various image processing algorithms, including object classification [10], background modeling [11, 12], and saliency detection [13, 14]. However, a major drawback of these methods is that their assumption of approximating the sea edge by a straight line is often violated in coastal waters and narrow marinas [5], which leads to poor obstacle detection. To avoid

✉ Hengyu Li
lihengyu@shu.edu.cn

Extended author information available on the last page of the article.

this problem, recent works [1–5] built upon probabilistic graphical models and proposed semantic-segmentation-based obstacle detectors for USVs, which have achieved excellent results without the help of the horizon and are the current state-of-the-art methods in this field. However, these approaches perform semantic segmentation on a reduced-size image to accelerate inference, which causes difficulties in small obstacle segmentation. In addition, these methods model the semantic structure of marine scenes using just simple features (e.g., pixel's positon and color values), which cannot address the environmental appearance diversity in the presence of scenery reflections, boat wakes, and visual ambiguities, leading to poor segmentation results.

In recent years, numerous semantic segmentation methods based on deep convolutional neural networks (DCNNs) [15–22] have been proposed for the scene perception of unmanned ground vehicles (UGVs). Due to the powerful feature extraction and representation capability of DCNNs, these methods significantly surpass conventional hand-crafted feature-based methods [17]. An increasing number of semantic segmentation networks have been employed in autonomous driving and achieved promising results. However, existing semantic segmentation networks proposed in UGVs cannot be readily applied for USVs because of many differences between the USV domain and the UGV domain. The most prominent difference is that the navigable surface of USVs is nonflat and dynamic and its appearance is affected by weather conditions compared with that of UGVs. Moreover, seawater is a reflective surface and the light from the surroundings may produce complex reflections on the water. In two recent studies [23, 24], several semantic segmentation networks that are widely used in the UGV domain have been evaluated on a challenging maritime dataset (MaSTr1325), including UNet [25], PSPNet [20] and DeepLabv2 [26]. The evaluation results have shown that these networks perform better than the conventional model-based methods [1, 2], but their performance is still susceptible to water reflections and visual ambiguities. Furthermore, most of them have difficulty in segmenting small-size obstacles. This is because small obstacles contain fewer pixels and the pooling layers in these networks may remove the features of small objects. In addition, many state-of-the-art semantic segmentation networks improve their performance by increasing the number of convolutional layers [17]. As a result, these networks require more memory to store redundant parameters and more time for inference, which are inappropriate for real-time obstacle map estimation of USVs.

Although some progress has been made in marine scene perception, obstacle map estimation based on computer vision for USVs still faces several difficulties due to dynamic and unconstrained environments. (i) Segmenting small obstacles is notoriously challenging because of their low resolution and difficulty in distinguishing them from sea surface disturbances, such as waves, wakes and sea foam. (ii) Irregular scenery reflection on the navigable surface of USVs is visually similar to its surrounding object, which increases the difficulty of water segmentation and obstacle detection. (iii) When the onboard camera is directly facing the sun, the illumination variation and sun glitter substantially reduce the performance of obstacle map estimation using color features. (iv) Adverse weather, such as fog and haze, can significantly degrade the visibility of marine scenes, which hinders the estimation of obstacle maps for USVs. Figure 1 shows various challenges in estimating obstacle maps for USVs.

To address these issues, we consider a lightweight encoder-decoder network (e.g., ENet [16]) as the backbone and propose a novel multistage feature aggregation and semantic feature separation network (FASNet) for real-time obstacle map estimation of USVs (see Fig. 2). More

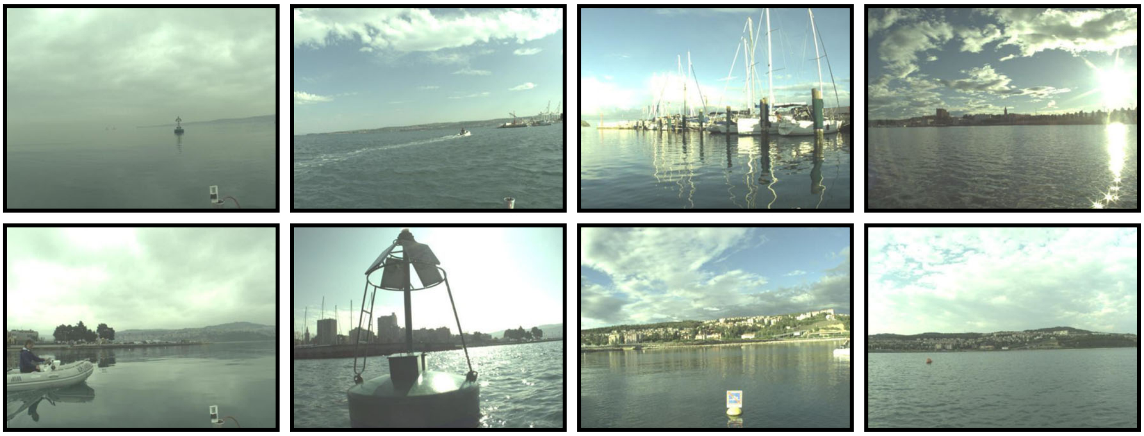


Fig. 1 Visual diversity in MaSTr1325 [23] ranging from foggy, boat wake, water reflection, sun glitter (top) and various obstacles (bottom)

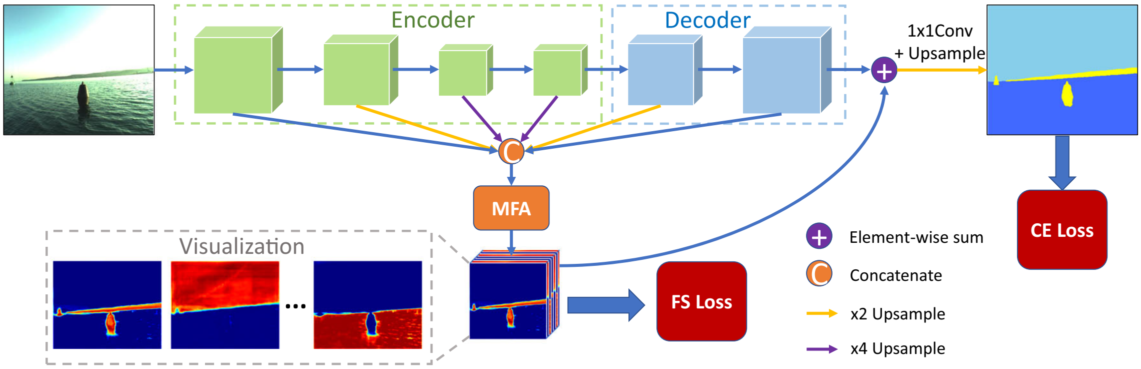


Fig. 2 Architecture of our proposed FASNet. Different stage features from the encoder and decoder of a lightweight backbone network (e.g., ENet [16]) are concatenated as the input of the multistage feature aggregation (MFA) module. In the MFA module, multiscale spatial and depthwise separable convolutions are applied to capture and fuse different scale features and context information. A semantic feature separation (FS) loss computed at the output of the MFA module enforces the learning of discriminative features, which are further used to guide the pixelwise prediction. A cross-entropy (CE) loss is also employed in the final pixelwise prediction to train the network

specifically, an efficient multistage feature aggregation (MFA) module is proposed to fuse the shallow and deep features of the backbone at different stages, which aims to improve the segmentation performance of small obstacles. To reduce the number of parameters and computational complexity, multiscale spatial and depthwise separable convolutions are applied to capture and fuse different-scale context information in the MFA module. In addition, we design a new loss function to enforce the separation among different semantic features, which provides a robust feature representation in the scenes with environmental reflections, salient wakes, and visual ambiguities. We evaluated the proposed FASNet on a challenging maritime segmentation dataset: MaSTr1325 [23]. The evaluation results showed that the proposed method achieved better performance than other state-of-the-art lightweight models. We also validated that various lightweight semantic segmentation networks achieved consistent performance improvement when our MFA module and loss function were adopted. The main contributions of this work include the following:

1) We proposed an efficient multistage feature aggregation (MFA) module to capture and fuse different scale features and context information of the backbone at different stages.
2) We designed a new loss function to increase the separation among different semantic features, achieving a more robust feature representation in highly diverse marine environments.
3) We presented a novel multistage feature aggregation and semantic feature separation network (FASNet) for real-time obstacle map estimation of USVs. In addition, we reported extensive quantitative and qualitative evaluation results of the proposed model on a public challenging segmentation dataset (MaSTr1325).

2 Related Works

Compared with the field of UGVs, camera-based environmental perception in USVs is still a relatively young research area. Since the obstacles that pose a threat to USVs are generally located below the horizon line, most conventional methods estimate a horizon line and then detect obstacles under the estimated horizon line using various object detection models (e.g., object classification [10], background modeling [11, 12], and saliency estimation [13, 14]). However, their common assumption that the horizon is approximated as a straight line is often violated in coastal waters or narrow marinas, which causes poor performance of obstacle detection. In addition, these methods that use bounding boxes to locate objects cannot provide accurate navigable water areas for collision avoidance in USVs. Moreover, the approaches proposed in [10, 11, 13, 14] perform obstacle detection with hand-crafted features, which poses difficulty in addressing the appearance diversity of a marine scene with water reflections, wakes, sun glitter or visual ambiguities.

Several recent works [1–5] have explored semantic segmentation methods based on a probabilistic graphical model to obtain obstacle maps for small-sized USVs without the guidance of the horizon line. In [1], a novel graphical model was proposed to account for the semantic structure of marine images observed from USVs. The model is optimized by an expectation maximization (EM) algorithm to fit each input image and generate an obstacle map. This approach has achieved excellent results in monocular perception of USVs, which can detect not only obstacles that stand out from the water surface but also floating flat obstacles. Bovcon et al. [2] extended the graphical model [1] with inertial measurement information to improve the segmentation in the presence of visual ambiguities. In [3, 4], the performance of the graphical model proposed in [1] has been enhanced by fusing inertial measurement data and stereo vision information. Liu et al. [5] designed a prior estimation network to enhance the performance of the graphical model [1]. These methods significantly outperform other related algorithms with respect to the obstacle perception of USVs and have achieved state-of-the-art performance. However, these methods still have drawbacks in the presence of sun glitter, reflections, and visual ambiguities because these approaches model the semantic structure of the marine scene using simple color and position features, which fail to fully capture the scene appearance diversity.

Due to the powerful feature extraction and representation capabilities of DCNNs, many researchers have developed various semantic segmentation methods based on DCNNs for the scene perception of UGVs [15, 17]. These methods significantly outperform conventional computer visual methods on many benchmarks [27, 28]. However, most state-of-the-art networks that achieve excellent performance are based on a very deep convolutional backbone (e.g., ResNet-101 [29]), such as PSPNet [20], DeepLabv3+ [21], and EMANet [22], which require a large amount of memory to store parameters and significant time for inference. Therefore, these methods can hardly execute in real time and are not appropriate for obstacle map estimation in USVs. For the purpose of real-time applications on mobile platforms with limited computing power, some efficient semantic segmentation networks are proposed, including ENet [16], FSSNet [17], ERFNet [18] and ESNet [19]. These lightweight networks exploit various solutions (e.g., depthwise separable convolution [30], atrous convolution [26] and factorized block [17]) to reduce running time. However, these networks exhibit slightly lower accuracy than other state-of-the-art deeper networks. In addition, existing semantic segmentation networks in UGVs cannot be readily used for USVs because of the significant environmental differences. For example, the most obvious difference is that the navigable surface of USVs is dynamic and affected by weather conditions. Two recent separate studies [23, 31] have tested the performances of the semantic segmentation networks that are widely used in UGVs on several marine datasets. Both of them have reached a consistent conclusion that commonly used semantic segmentation networks from UGVs are susceptible to water reflections and visual ambiguities. Furthermore, most of these networks are prone to misclassifying small obstacles as background regions.

3 Method

The overall architecture of FASNet is introduced in Section 3.1. Section 3.2 details an efficient multistage feature aggregation (MFA) module, and a novel semantic feature separation (FS) loss is described in Section 3.3.

3.1 Overall Architecture

The overall architecture of our FASNet is shown in Fig. 2. In consideration of the limitations of USVs with respect to computing power, we choose an off-the-shelf efficient semantic segmentation network (e.g., ENet [16]) as the backbone of our network. In practice, the backbone network can be replaced by other lightweight networks with an encoder-decoder structure, such as FSSNet [17], ERFNet [18] and ESNet [19]. The encoder in the backbone network spatially reduces the feature maps and captures high-level semantic information, while the decoder gradually recovers the detailed object boundaries and upsamples the feature maps from the encoder to the same size as the original image.

In the marine scenes observed from USVs, various small obstacles may appear on the preplanned path and threaten the safe navigation of USVs. However, existing semantic segmentation networks exhibit an unfavorable performance in small obstacle segmentation. The backbone network computes a feature hierarchy layer by layer. The convolutional blocks at the different stages of the backbone produce feature maps with different spatial resolutions, which inherently contain multiscale context information [32]. To improve the performance in small obstacle segmentation, we propose an efficient multistage feature aggregation (MFA) module to capture and aggregate multiscale features from the multistage outputs of the backbone, and then the feature map computed by the MFA module is used to guide the final pixelwise prediction. In addition, a novel semantic feature separation (FS) loss computed at the output of the MFA module enforces learning of discriminative features, which leads to a robust feature representation in the presence of visual ambiguities, salient reflections or wakes. In the final pixelwise prediction, a cross-entropy (CE) loss is also employed to train the network.

3.2 Multistage Feature Aggregation (MFA) Module

For semantic segmentation of marine images, visual information needs to be processed on various scales. As discussed in Section 3.1, the feature maps of a deep convolutional network at different stages inherently contain multiscale context information. Empirically, combining low-level feature maps with high-level feature maps helps improve the segmentation performance of small objects. Therefore, we design an MFA module to fuse the different stage feature maps of the backbone network. To reduce the amount of parameters and computational complexity, multiscale spatial and depthwise separable convolutions are applied to process the features on various scales in the MFA module.

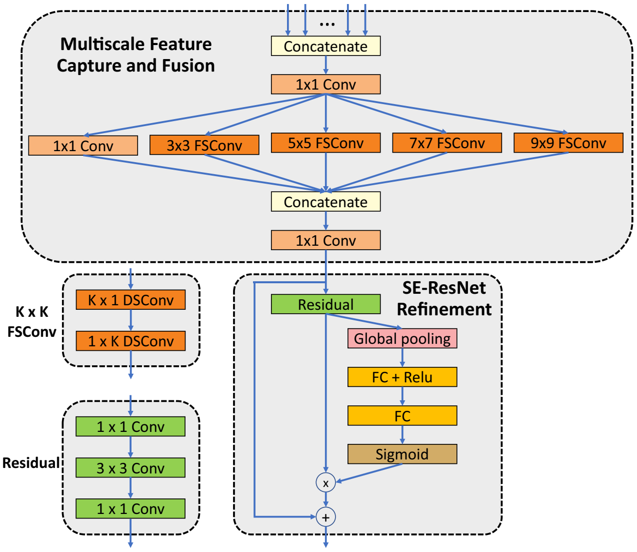


Fig. 3 Multistage feature aggregation (MFA) module. Note that each convolution is followed by a batch normalization layer [33] and rectified linear unit (ReLU) function [34]. *Conv*: standard convolution, *DSConv*: depthwise separable convolution, *FSConv*: fully separable convolution (spatial and depthwise separable convolution), *FC*: fully connected layer, *K*: the filter size of a convolution. *Residual*: the convolutions in the standard residual bottleneck

Figure 3 shows the framework of the MFA module, consisting of two stages: (i) multiscale feature capture and fusion, and (ii) SE-ResNet refinement. In the first stage of the MFA module, the multistage feature maps from the backbone network are resized and concatenated as a single feature map. The channel depth of the concatenated feature map is reduced by a 1 × 1 convolution. To better capture the structure and context information of the image, inspired by [35], five parallel convolutions with different scales (e.g., 1 × 1, 3 × 3, 5 × 5, 7 × 7, and 9 × 9) are performed on the feature map. Different from the method in [35] that adopts several parallel standard convolutions, we spatially decompose the standard convolution into two single dimensional convolutions to reduce the number of parameters and running time. For example, a $k \times k$ standard convolution can be replaced by a sequence of $k \times 1$ and $1 \times k$ convolutions, which is termed spatial separable convolution. As a result, this replacement makes the parameters $\frac{k}{2}$ times less than those of the standard convolution, while maintaining the equal size of the receptive field. In addition, similar to [36], depthwise separable convolution [30] is also introduced in each spatial separable convolution to further reduce the computation and parameters. This separable convolution, which considers both the spatial dimensions and the depth dimensions, is referred to as fully separable convolution in this paper (see Fig. 4). Following the parallel multiscale convolutions, a concatenation operation is applied to merge the output feature maps, and a 1 × 1 dimension reducing convolution is employed to reduce the depth of the concatenated feature map. Here, the context information extraction and fusion performed by the parallel multiscale convolutions and the dimension reducing convolution can be described as

$$\boldsymbol{X}^{out} = \delta \left(\boldsymbol{W} * \bigcup_{i=1}^{m} \delta_i \left(\boldsymbol{W}_i * \boldsymbol{X}^{in} + \boldsymbol{b}_i \right) + \boldsymbol{b} \right) , \qquad (1)$$

where $\boldsymbol{X}^{in}$ denotes the input feature map, $\boldsymbol{W}_i$ and $\boldsymbol{b}_i$ are the weight and bias of the i-th scale fully separable convolution or standard convolution, respectively, and m is the number of different scale convolutions. $\boldsymbol{W}$ and $\boldsymbol{b}$ are the weight and bias, respectively, of the dimension reducing convolution with kernel size 1 × 1. δ denotes the rectified linear unit (ReLU) activation function [34]. $*$ and $\bigcup$ represent the convolution operation and the concatenation operation, respectively.

To better refine the feature map obtained by the parallel multiscale convolutions, we also add a SE-ResNet module [37] as the final block of our MFA module. The SE-ResNet module builds upon a standard residual bottleneck [29] and inserts a spatial squeeze and channel excitation block to emphasize attention on the selected channels of the feature map. The output feature map of the MFA module is added with the feature map of the decoder to guide the pixelwise prediction, as illustrated in Fig. 2.

3.3 Enhancing Semantic Features Separation

In coastal waters or narrow marinas, the surrounding environment and sunlight may be reflected on the dynamic water surface, which causes significant changes in the appearance of the water. In addition, the appearance of the marine scene is greatly affected by unfavorable weather conditions. In particular, haze or foggy weather will cause visual ambiguities and make the boundary between sea

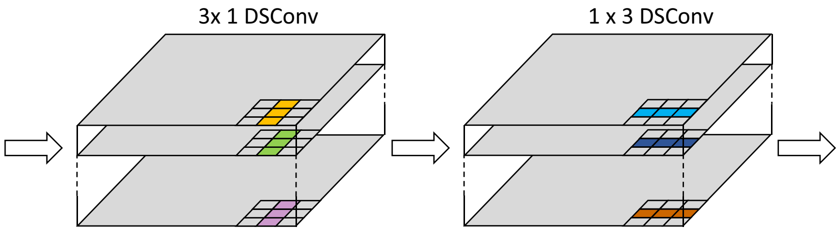


Fig. 4 3 × 3 fully separable convolution decomposes a 3 × 3 standard convolution in the spatial and depth dimensions. *DSConv*: depthwise separable convolution

and sky unclear. Due to these challenges in the marine environment, existing semantic segmentation networks still encounter difficulties in the obstacle map estimation of USVs [23]. To enhance the robustness of our network in the abovementioned challenging scenarios, we design a novel loss function to enforce semantic feature separation, which is applied at the output of the MFA module.

Denote $\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$ as the feature map computed by the MFA module, where $H \times W$ means the spatial dimension. To facilitate the description, we reshape $\boldsymbol{X}$ into $N \times C$, where $N = H \times W$. Further, $\boldsymbol{X}$ can be expressed as $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1:N}$, where $\boldsymbol{x}_i \in \mathbb{R}^C$ represents the C dimensional feature vector at pixel i. To estimate the obstacle map for USVs, networks need to classify each pixel on an image into three categories: obstacle, water and sky. Thus, $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1:N}$ can be divided into $\boldsymbol{X}^{obs} = \{\boldsymbol{x}_i\}_{i\in Obstacle}$, $\boldsymbol{X}^{wat} = \{\boldsymbol{x}_i\}_{i\in Water}$ and $\boldsymbol{X}^{sky} = \{\boldsymbol{x}_i\}_{i\in Sky}$ according to the annotated labels $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1:N}$, where $\boldsymbol{y}_i$ is the annotated label of pixel i. To encourage different semantic features to cluster towards their centroids, we assume a Gaussian mixture model to fit their distributions:

$$p(\boldsymbol{x}_i|\Theta) = \sum_{k=1}^{3} \alpha_k \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \, , \tag{2}$$

where $\phi(\cdot|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a Gaussian kernel. $\Theta = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1:3}$ denotes the means and covariances of the three Gaussian components, which can be computed by $\boldsymbol{X}^{obs}$, $\boldsymbol{X}^{wat}$ and $\boldsymbol{X}^{sky}$. α_k is the weight for the k-th Gaussian component, meeting $0 \le \alpha_k \le 1$ and $\sum_{k=1}^{3} \alpha_k = 1$. A Gaussian mixture model (GMM) is a probabilistic model that assumes all the observed data points are generated from a mixture of a finite number of Gaussian distributions, which is parameterized by the mixture component weights and the component Gaussian parameters [22, 38]. GMM is an unsupervised learning method for data clustering and data mining. In this part, GMM is used to fit the learned features and cluster different semantic features towards their centroids.

Using Bayes' theorem and the model parameters Θ and α_k, the posterior probability z_{ik} that a data point $\boldsymbol{x}_i$ belongs to the k-th Gaussian component can be calculated as follows:

$$\begin{aligned} z_{ik} &= p(\gamma_i = k|\boldsymbol{X}, \Theta) \\ &= \frac{p(\gamma_i = k, \boldsymbol{x}_i|\Theta)}{\sum_{l=1}^{3} p(\gamma_i = l, \boldsymbol{x}_i|\Theta)} \\ &= \frac{p(\boldsymbol{x}_i|\gamma_i = k, \Theta)p(\gamma_i = k|\Theta)}{\sum_{l=1}^{3} p(\boldsymbol{x}_i|\gamma_i = l, \Theta)p(\gamma_i = l|\Theta)} \\ &= \frac{\alpha_k \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{3} \alpha_l \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \, , \end{aligned} \tag{3}$$

where γ_i denotes the label of the i-th data point $\boldsymbol{x}_i$, which is a latent random variable and holds $p(\gamma_i = k) = \alpha_k$.

The posterior z_{ik} is a measure of the distance or similarity between the feature $\boldsymbol{x}_i$ and the k-th Gaussian component. To facilitate the implementation of $\phi(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ in Eq. 3, we omit the $\boldsymbol{\Sigma}_k$ and simply adopt the exponential inner dot $\exp(\boldsymbol{x}_i^T \boldsymbol{\mu}_k)$ to replace the Gaussian kernel. In addition, we also assume that any α_k is equal to $1/3$. Thus, Eq. 3 can be implemented as a matrix multiplication plus one softmax layer:

$$\boldsymbol{Z} = \text{softmax}\left(\frac{\boldsymbol{X}}{||\boldsymbol{X}||_2} \cdot (\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||_2})^T\right) \, , \tag{4}$$

where $\boldsymbol{X} \in \mathbb{R}^{N\times C}$ is the input feature map, and $\boldsymbol{\mu} \in \mathbb{R}^{3\times C}$ is the mean matrix which consists of the means of three Gaussian components. $||\cdot||_2$ is the L2 normalization for stability.

By combining (4) and conventional cross-entropy loss, we propose a novel semantic feature separation loss function to enforce discriminative representation learning:

$$L_{fs} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{3} y_{ik} \log z_{ik} \, , \tag{5}$$

where z_{ik} and y_{ik} refer to the estimated posterior probability and its corresponding ground truth for pixel i. N is the number of pixels on the feature map X computed by the MFA module.

For the final semantic segmentation prediction, we employ a weighted cross-entropy loss L_{ce} to train our network as follows:

$$L_{ce} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{3} w_k q_{ik} \log p_{ik} \, , \tag{6}$$

where p_{ik} and q_{ik} represent the prediction result of FASNet and its corresponding ground truth, respectively, for pixel i. M is the number of pixels in the output of FASNet. w_k denotes a class weight for improving FASNet on the imbalanced training dataset. According to [16, 17], we use the class weighting scheme:

$$w_k = \frac{1}{\log(c + \hat{p}_k)} \, , \tag{7}$$

where $\hat{p}_k$ represents the probability of a class, which is obtained by calculating the weight of each category in the training dataset. c is an additional hyperparameter. In this paper, we set c to 1.02 and control the class weights between 1 and 50.

Combining (5) and (6), the total loss of FASNet can be represented as follows:

$$L_{FASNet} = L_{ce} + L_{fs} \, . \tag{8}$$

4 Results

In this section, we evaluate the effectiveness of FASNet on a challenging marine dataset (MaSTr1325). The dataset and evaluation metrics are described in Section 4.1. The implementation details are provided in Section 4.2. Ablation studies and a comparison with other state-of-the-art networks are given in Sections 4.3 and 4.4, respectively. In addition, we also compare FASNet with other state-of-the-art conventional methods in Section 4.5. In Section 4.6, we apply our MFA module and FS loss function to existing lightweight models for wide applicability evaluation.

4.1 Dataset and Evaluation Metrics

The performance of our FASNet was evaluated on a challenging marine semantic segmentation dataset (MaSTr1325) [23], which consists of 1,325 fully annotated images with a resolution of 512 × 384 pixels. These images were captured by a small-sized USV over a two-year span, covering a range of realistic conditions encountered in a coastal surveillance task. In particular, attention was paid to include various weather conditions and times of day to ensure the variety of the captured dataset. Each image was manually annotated by human annotators with three categories (obstacle, water and sky). Some examples are shown in Fig. 1. To train and evaluate semantic segmentation networks, we randomly divided the dataset into 905 images for training, 120 images for validation, and 300 images for testing. Similar to [23], we applied central rotations of $\pm\{5, 15\}$ degrees on each image (see Fig. 5) to increase the complexity of the dataset. We refer to this expanded dataset as augmented-MaSTr1325.

To quantitatively evaluate the accuracy of our proposed FASNet and other state-of-the-art deep segmentation networks, several conventional semantic segmentation evaluation metrics were chosen, including per-class intersection over union (per-class IoU), mean intersection over union (mIoU) and frequency weighted intersection over union (FWIoU) [39]. In the quantitative comparison with other state-of-the-art conventional methods on the task of water segmentation, we adopted classical pixel-based metrics [40] to evaluate the segmentation performance, including precision, recall and F1-measure. In addition, we also employed the number of parameters, the number of floating point operations (FLOPs) and the average execution time per frame (ΔT) to measure the implementation efficiency.

4.2 Implementation Details

All semantic segmentation networks in the evaluation were implemented in PyTorch and trained end-to-end. For real-time applications, we adopted lightweight ENet [16] as the backbone of FASNet. The backbone and other comparison networks were pretrained on Cityscapes [27]. The convolutional layers in the MFA module were randomly initialized using the Kaiming method [34]. We used a learning rate of 0.001 and weight decay of 0.0004. An Adam optimizer was utilized to iteratively update network weights in the training with 150 epochs, and the learning rate was decayed by a factor of 0.2 every 30 epochs. The size of a mini-batch was set to 8. To avoid overfitting, typical data augmentations in semantic segmentation networks were used, including color jittering and random horizontal flipping. All networks were trained and tested on an industrial computer with an Intel Core i7-6700K 4 GHz

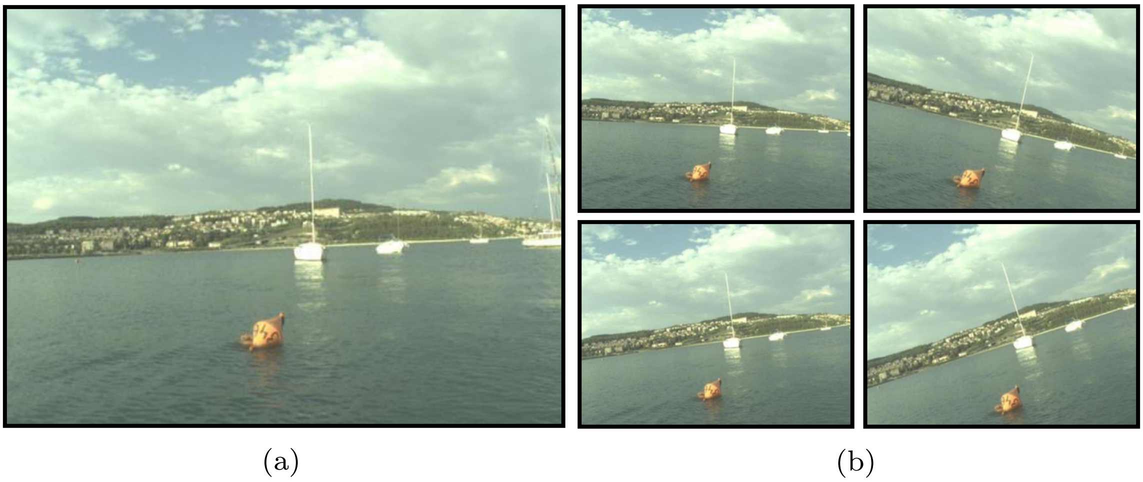


Fig. 5 An example of data augmentation using central rotation. **a** Input image. **b** Rotation augmented variations

Table 1 Ablation study results on the MaSTr1325 test dataset

Network	Backbone	MFA	FS	Per-class IoU (%)			mIoU (%)
				Obstacle	Water	Sky	
ENet [16]	–			90.02	98.88	98.96	95.95
FASNet^{-}	ENet [16]	✓		90.63	98.98	99.01	96.21
FASNet	ENet [16]	✓	✓	**91.77**	**99.06**	**99.17**	**96.67**

The best results are highlighted in **bold**

CPU with 32 GB RAM and an externally connected NVIDIA GTX1080Ti GPU with 11 GB of memory.

4.3 Ablation Studies

There are two major novelties in our FASNet: the MFA module and the feature separation (FS) loss. To explore how each novelty contributes to enhancing the segmentation performance, we conducted ablation studies on the original MaSTr1325 dataset, which was randomly divided into 905 images for training, 120 images for validation, and 300 images for testing. Table 1 shows the ablation study results of FASNet with different settings, in which FASNet with the FS loss removed is termed FASNet^{-}. As shown in Table 1, both the MFA and FS loss boosted the segmentation performance. Compared with the baseline ENet, the proposed FASNet, which contains the MFA module and FS loss, achieved a 0.72% mIoU and 1.75% obstacle-class IoU increase. To better understand the effect of FS loss, we visualized the feature maps computed by the MFA module in FASNet^{-} and FASNet and selected the most activated feature map channels for comparison in Fig. 6. As shown in Fig. 6, FS loss can significantly enhance the sharpness of the boundary and the completeness of different categories, which enables a more discriminative feature representation.

4.4 Comparison with state-of-the-art deep segmentation networks

To validate the effectiveness of FASNet, we compared it with recent state-of-the-art segmentation networks, including ENet [16], FSSNet [17], ERFNet [18], ESNet [19], PSPNet [20] and DeepLabv3+ [21]. Considering the limited computing power of mobile platforms, PSPNet and DeepLabv3+ adopt ResNet-50 [29] as the backbone. For

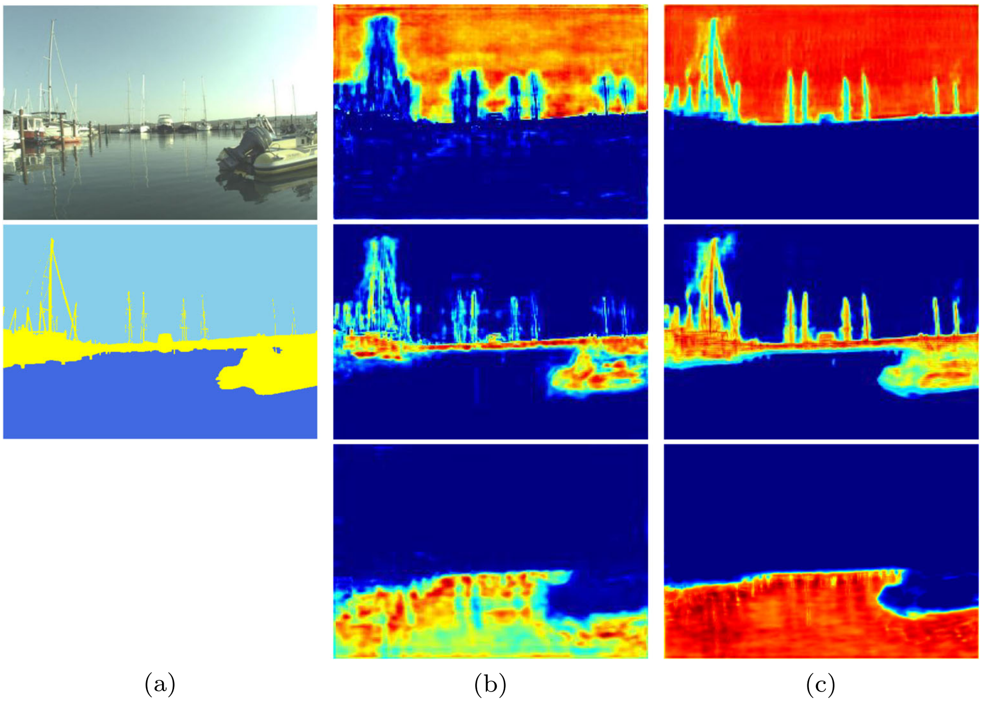


Fig. 6 Feature map visualization of the MFA module in FASNet and FASNet^{-}. (a) Input image and its ground truth. (b) Visualized results in FASNet^{-}. (c) Visualized results in FASNet. FASNet^{-} represents the variant of FASNet without the FS loss. In comparison, FASNet with the FS loss can learn a more discriminative feature representation in the complicated marine environment

Table 2 Quantitative segmentation comparison with various state-of-the-art networks on the augmented-MaSTr1325 test dataset

Network	Backbone	Per-class IoU (%)			mIoU (%)	FWIoU (%)
		Obstacle	Water	Sky		
ENet [16]	–	90.48	98.66	98.71	95.96	97.83
FSSNet [17]	–	90.25	98.59	98.70	95.85	97.77
ERFNet [18]	–	89.84	98.52	98.64	95.66	97.66
ESNet [19]	–	90.07	98.57	98.66	95.76	97.72
PSPNet [20]	ResNet-50 [29]	89.37	98.59	98.51	95.49	97.58
DeepLabv3+ [21]	ResNet-50 [29]	90.19	98.65	98.66	95.83	97.76
FASNet	ENet [16]	**92.24**	**98.88**	**99.00**	**96.71**	**98.24**

The best results are highlighted in **bold**

a fair comparison, all networks were trained and tested on the augmented-MaSTr1325 dataset using the same learning scheduler, as detailed in Sections 4.1 and 4.2.

The quantitative evaluation results of various networks on the augmented-MaSTr1325 test dataset are summarized in Table 2. As observed from Table 2, our FASNet achieved the highest accuracy for all 3 classes and achieved the best results of 96.71% mIoU and 98.24% PWIoU. In particular, FASNet significantly improved the segmentation performance for the obstacle category and provided 1.76% improvement compared with the second-best ENet. The results demonstrated that the proposed MFA module with the FS loss function can help the backbone network improve the segmentation accuracy of marine images, especially for obstacle regions.

Figure 7 displays representative test images, ground truths, and the segmentation results of different networks. The qualitative comparison revealed that FASNet was more robust than all other networks under various extreme conditions. As shown in the first column of Fig. 7, FASNet estimated the water region more accurately in the presence of severe environmental reflections, while the other methods produced false positives. From the second column of Fig. 7, it can be observed that the other networks performed poorly in the presence of visual blur, which renders the extent of the water region significantly oversegmented. The comparison in the third column of Fig. 7 shows that our FASNet was better than the other networks at addressing distinct wakes caused by boats. In these scenes, the other networks tended to misclassify the wake edges as obstacles. The performance of FASNet in these challenging scenes can be attributed to the discriminative feature representation trained by FS loss. FASNet can effectively increase the separation among different semantic features, which enhances the discriminative ability of FASNet in the presence of environmental reflections, salient wakes, and visual ambiguities. As observed from the last column of Fig. 7, our method accurately segmented the small obstacle from the water region, while several state-of-the-art models performed poorly. There are two major reasons for the excellent performance of FASNet in small obstacle segmentation. The first reason is that the MFA module in the network can help capture and fuse the context information of different scale obstacles. The second reason is that the FS loss can enhance the discrimination between small obstacles and the background, which improves the performance improvement of FASNet.

To evaluate the computational performance of FASNet, we further compared it with other segmentation networks in terms of model complexity and execution time (see Table 3). In deep learning, network complexity usually involves space complexity and time complexity. The space complexity of a deep learning network refers to the number of parameters in the network, which is directly proportional to the amount of memory consumed by the network. Similarly, the time complexity, which is also referred to as computational complexity, refers to the number of float-point operations (FLOPs) that the network needs to perform inference on a single sample. As shown in Table 3, the parameters of our FASNet were fewer than the lightweight ERFNet and ESNet. Although they had comparable FLOPs, our FASNet was much faster than ERFNet and ESNet. The discrepancy between the FLOPs and the execution time is attributable to the notion that the inference time of a network is also affected by the memory access cost and degree of parallelism. Compared with the baseline ENet, FASNet benefited from the application of the fully separable convolution, and therefore, only increased very few parameters and FLOPs. The quantitative results (Table 3 last column) showed that our proposed FASNet achieved an average execution speed of 58 frames per second (fps) on an NVIDIA GTX1080Ti GPU. Although the speed of FASNet was slower than that of the baseline ENet, it did not slow down the overall onboard control system in practice because the cameras for the MaSTr1325 [23] were limited to 10 fps.

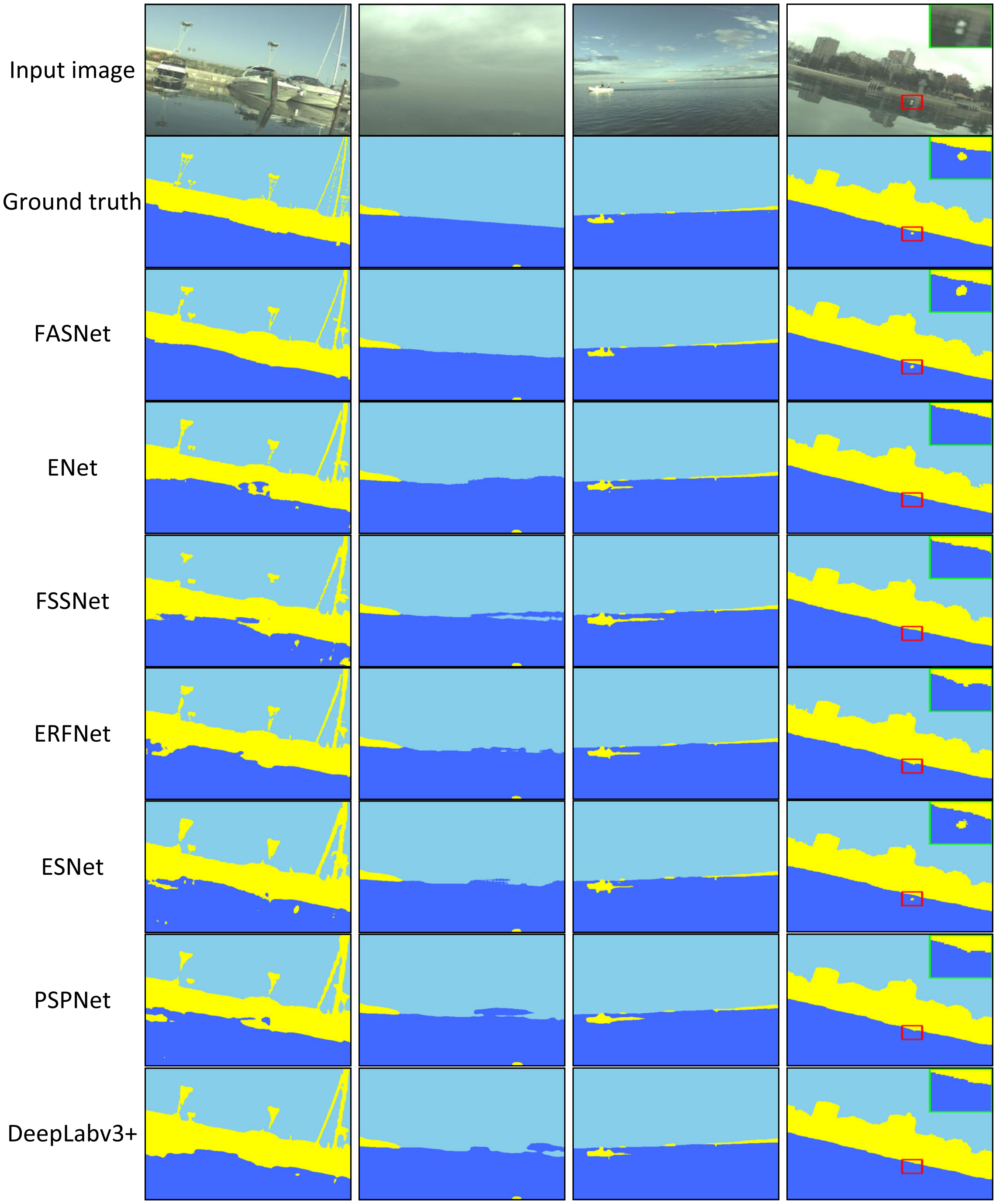


Fig. 7 Qualitative examples of various semantic segmentation networks on the augmented-MaSTr1325 test dataset. It can be seen that FASNet outperforms other networks in the presence of severe environmental reflection (see the first column), visual ambiguity (see the second column), salient wake (see the third column) and small obstacle (see the last column)

4.5 Comparison with state-of-the-art conventional methods

We also compared the proposed FASNet with current state-of-the-art conventional methods: SSM [1] and PEN-SSM [5]. SSM and PEN-SSM both assume a probabilistic graphical model, which is fitted to each image by iterative expectation maximization steps and simultaneously classifies pixels into water and non-water according to the corresponding posterior distributions. Thus, an online obstacle map for USV can be obtained. To facilitate comparison, we only evaluated the accuracy of these methods for the task of water segmentation.

Table 3 Comparison of network complexity and average execution time on the augmented-MaSTr1325 test dataset

Network	Backbone	Parameters (M)	FLOPs (G)	$\Delta T\,(ms)$
ENet [16]	–	0.37	1.63	10
FSSNet [17]	–	**0.20**	**1.59**	**7**
ERFNet [18]	–	2.06	11.11	27
ESNet [19]	–	1.66	10.15	30
PSPNet [20]	ResNet-50 [29]	48.77	138.56	249
DeepLabv3+ [21]	ResNet-50 [29]	42.53	142.66	491
FASNet	ENet [16]	0.61	14.21	17

The best results are highlighted in **bold**

The quantitative results of the observed methods for water segmentation are summarized in Table 4. Although our proposed FASNet was slightly slower than SSM, it achieved the highest precision and recall values, along with the highest F1-measure. In the F1-measure, FASNet significantly outperformed SSM and PEN-SSM by 10.14% and 4.46%, respectively. The quantitative comparison results showed that our method was more reliable than SSM and PEN-SSM on the task of water segmentation. This was because the proposed FASNet can effectively learn more discriminative features from the annotated data, thereby improving the accuracy of water segmentation. Figure 8 shows a further qualitative comparison of the methods. Note that SSM and PEN-SSM both performed poorly in the presence of scenery reflections, visual ambiguities, sun glitter, and obstacles visually similar to water (see Fig. 8, first four rows). The reason for the poor performance was that they relied on pixel's position and color features, which failed to correctly address the diversity of a complex marine scene, resulting in poor water segmentation. In addition, SSM and PEN-SSM performed water segmentation on a reduced-size image to accelerate the inference process, which caused difficulty in small obstacle segmentation (see Fig. 8 last row). In comparison, our FASNet performed well in these challenging scenes, which was reflected in the quantitative comparison results.

4.6 Wide Applicability Evaluation

Embedding a backbone trained on related tasks into the semantic segmentation network for obstacle map estimation in marine scenes can accelerate the training process and improve the segmentation performance. Different backbones in the same network architecture may show different segmentation performance and efficiency. In practice, we usually select the most appropriate one from the existing backbones according to the performance of the computing device. To adjust the selected backbone to obstacle map estimation for USVs, it requires a general technique to help the backbone improve the segmentation performance while running in real-time. Our MFA module together with the FS loss function is an efficient plug-and-play module, which is readily applied to existing lightweight backbones and renders them more suitable for maritime obstacle map estimation.

To evaluate the wide applicability of our MFA module with the FS loss function, we replaced the backbone in FASNet with various lightweight models, including FSSNet [17], ERFNet [18], and ESNet [19]. All networks in the wide applicability evaluation adopted the same learning scheduler, as detailed in Section 4.2. The training and testing of the networks were performed on the augmented-MaSTr1325 dataset. The evaluation results, which are

Table 4 Quantitative comparison with state-of-the-art conventional methods in terms of water segmentation on the augmented-MaSTr1325 test dataset

Method	Backbone	$\Delta T\,(ms)$	Precision (%)	Recall (%)	F1-measure (%)
SSM [1]	–	**11**	82.77	96.82	89.25
PEN-SSM [5]	–	21	93.88	96.01	94.93
FASNet	ENet [16]	17	**99.62**	**99.17**	**99.39**

The best results are highlighted in **bold**

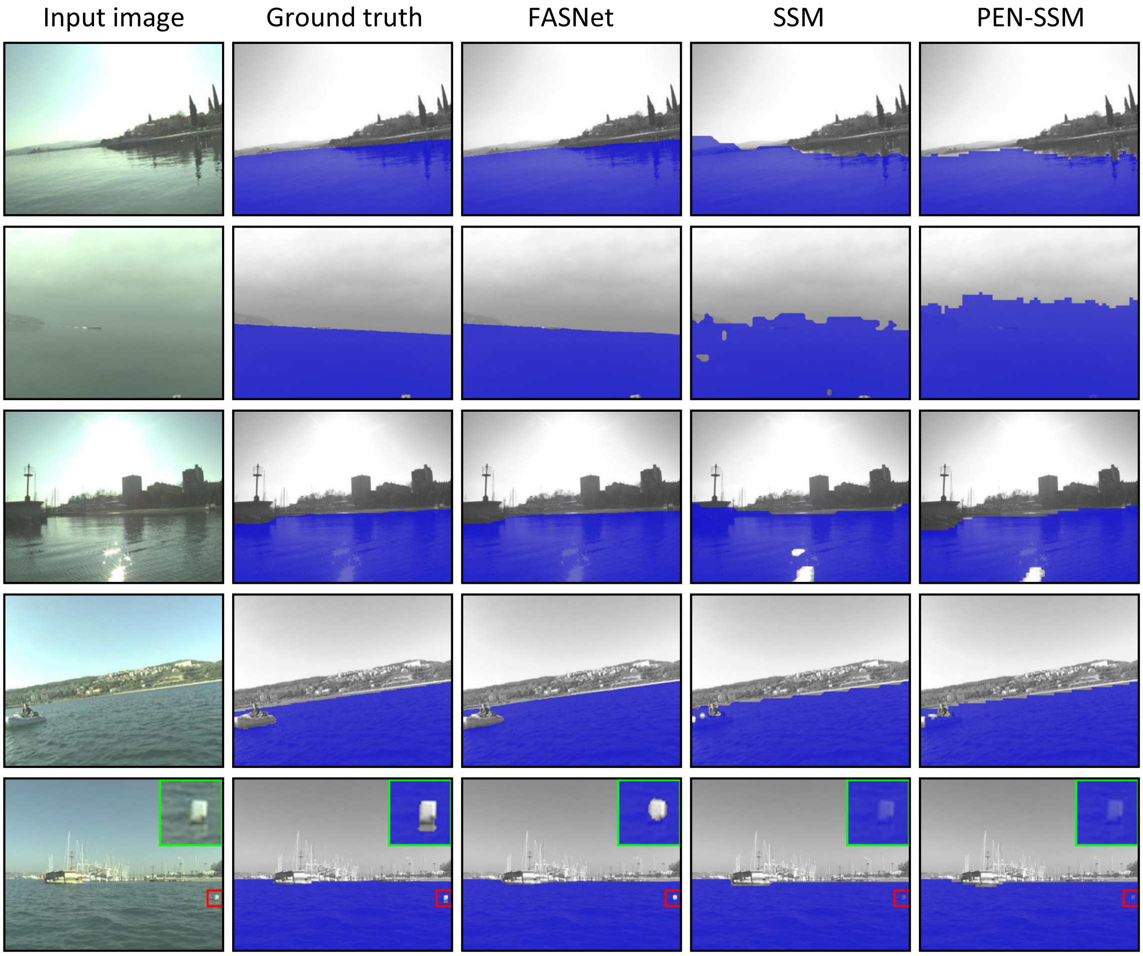


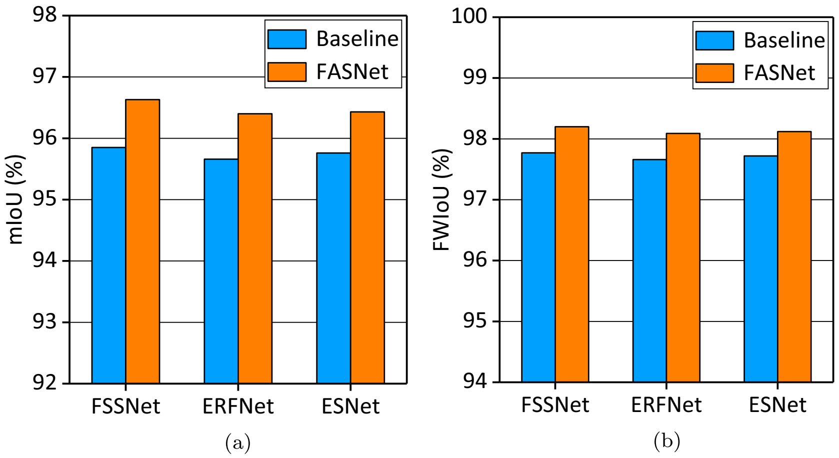
Fig. 8 Qualitative comparison of SSM [1], PEN-SSM [5] and our proposed FASNet on the task of water segmentation. The detected water region is denoted in blue. It can be observed that FASNet is superior to the conventional methods (SSM and PEN) in the complex scenes with scenery reflections (see the first row), visual ambiguities (see the second row), sun glitter (see the third row), obstacles visually similar to water (see the fourth row), and small obstacle (see the last row)

shown in Fig. 9, indicate that all networks with the FASNet framework consistently outperformed the corresponding baseline networks with significant increases in terms of mIoU and PWIoU.

5 Conslusion

This paper presented a novel semantic segmentation network, FASNet, for real-time obstacle map estimation of

Fig. 9 Comparison of **a** mIoU and **b** FWIoU between various baselines and corresponding FASNets on the augmented-MaSTr1325 test dataset. In comparison, all baseline networks achieve a consistent performance increase when using the FASNet framework



USVs. The network was evaluated on a public marine scene dataset (MaSTr1325), which demonstrated that FASNet outperformed existing state-of-the-art lightweight networks and achieved 96.71% mIoU, while running over 58 fps. In particular, FASNet significantly improved the segmentation performance for the obstacle category and provided 1.76% higher obstacle-class IoU than the second-best network (ENet [16]). The applicability evaluation also proved the effectiveness of the proposed MFA module and FS loss. Our future work will focus on exploring obstacle map estimation based on multisensor fusion to expand the application of USVs at night, and apply the estimation results as feedback for collision avoidance.

Acknowledgements The authors thank Xiaokang Yang and Rui Zhang for their assistance in the experimental evaluation. The authors also gratefully acknowledge the helpful comments and suggestions of the editor and anonymous reviewers, which have improved the presentation.

Author Contributions J. L. (Jingyi Liu) and H. L. conceived the idea. J. L. (Jingyi Liu), H. L. and J. L. (Jun Luo) designed the experiments. J. L. (Jingyi Liu) carried out programming, adjustment and data analysis. J. L. (Jingyi Liu) and Y. S. wrote the manuscript. All authors reviewed the final manuscript.

Funding This work was supported in part by the National Key Research and Development Program of China (grant number 2018YFB1304503), the National Natural Science Foundation of China (grant numbers 6193308 and 61625304) and the Shanghai Natural Science Foundation (grant number 18ZR1415300).

Availability of data and material The MaSTr1325 dataset that was used to train and evaluate FASNet is made publicly available at https://vicos.si/Projects/Viamaro.

Code Availability A public version of FASNet is available at https://github.com/aluckyi/FASNet.

Declarations

Conflict of Interests The authors declare no conflict of interest.

References

1. Kristan, M., Kenk, V.S., Kovačič, S., Perš, J.: Fast image-based obstacle detection from unmanned surface vehicles. IEEE Trans. Cybern. **46**(3), 641–654 (2016)
2. Bovcon, B., Perš, J., Kristan, M., et al.: Improving vision-based obstacle detection on USV using inertial sensor. In: Image and Signal Processing and Analysis (ISPA), 2017 10th International Symposium on, pp. 1–6. IEEE (2017)
3. Bovcon, B., Mandeljc, R., Perš, J., Kristan, M.: Stereo obstacle detection for unmanned surface vehicles by IMU-assisted semantic segmentation. Robot. Auton. Syst. **104**, 1–13 (2018)
4. Bovcon, B., Kristan, M.: Obstacle detection for USVs by joint stereo-view semantic segmentation. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5807–5812. IEEE (2018)
5. Liu, J., Li, H., Luo, J., Xie, S., Sun, Y.: Efficient obstacle detection based on prior estimation network and spatially constrained mixture model for unmanned surface vehicles. J. Field Robot. **38**(2), 212–228 (2021)
6. Zhao, Y., Li, W., Shi, P.: A real-time collision avoidance learning system for Unmanned Surface Vessels. Neurocomputing **182**, 255–266 (2016)
7. Chen, J., Pan, W., Guo, Y., Huang, C., Wu, H.: An obstacle avoidance algorithm designed for USV based on single beam sonar and fuzzy control. In: 2013 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 2446–2451. IEEE (2013)
8. Onunka, C., Bright, G.: Autonomous marine craft navigation: On the study of radar obstacle detection. In: 2010 11th International Conference on Control Automation Robotics & Vision, pp. 567–572. IEEE (2010)
9. Peng, Y., Yang, Y., Cui, J., Li, X., Pu, H., Gu, J., Xie, S., Luo, J.: Development of the USV 'JingHai-I' and sea trials in the Southern Yellow Sea. Ocean engineering **131**, 186–196 (2017)
10. Bloisi, D.D., Previtali, F., Pennisi, A., Nardi, D., Fiorini, M.: Enhancing automatic maritime surveillance systems with visual information. IEEE Trans. Intell. Transp. Syst. **18**(4), 824–833 (2016)
11. Zhang, Y., Li, Q.Z., Zang, F.N.: Ship detection for visual maritime surveillance from non-stationary platforms. Ocean Eng. **141**, 53–63 (2017)
12. Mou, X., Wang, H.: Image-based maritime obstacle detection using global sparsity potentials. J. Inf. Commun. Converg. Eng. **14**(2), 129–135 (2016)
13. Wang, H., Wei, Z., Wang, S., Ow, C.S., Ho, K.T., Feng, B., Lubing, Z.: Real-time obstacle detection for unmanned surface vehicle. In: Defense Science Research Conference and Expo (DSR), vol. 2011, pp. 1–4. IEEE (2011)
14. Shi, J., Jin, J., Zhang, J.: Object detection based on saliency and sea-sky line for USV vision. In: 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), pp. 1581–1586. IEEE (2018)
15. Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., Dietmayer, K.: Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Trans. Intell. Transp. Syst. (2020)
16. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147 (2016)
17. Zhang, X., Chen, Z., Wu, Q.J., Cai, L., Lu, D., Li, X.: Fast Semantic Segmentation for Scene Perception. IEEE Trans. Ind. Inf. **15**(2), 1183–1192 (2019)
18. Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R.: Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Trans. Intell. Transp. Syst. **19**(1), 263–272 (2017)
19. Wang, Y., Zhou, Q., Xiong, J., Wu, X., Jin, X.: ESNet: An efficient symmetric network for real-time semantic segmentation. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pp. 41–52. Springer (2019)
20. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
21. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
22. Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H.: Expectation-maximization attention networks for semantic segmentation. In:

Proceedings of the IEEE International Conference on Computer Vision, pp. 9167–9176 (2019)

23. Bovcon, B., Muhovič, J., Perš, J., Kristan, M.: The MaSTr1325 dataset for training deep USV obstacle detection models. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3431–3438. IEEE (2019)
24. Bovcon, B., Kristan, M.: Benchmarking semantic segmentation methods for obstacle detection on a marine environment. In: 24th Computer Vision Winter Workshop, pp. 1–9 (2019)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
26. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)
27. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
28. Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: European Conference on Computer Vision, pp. 44–57. Springer (2008)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
30. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
31. Cane, T., Ferryman, J.: Evaluating deep semantic segmentation networks for object detection in maritime surveillance. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2018)
32. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
33. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167 (2015)
34. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
35. Fan, X., Yang, Y., Deng, C., Xu, J., Gao, X.: Compressed multi-scale feature fusion network for single image super-resolution. Signal Process. **146**, 50–60 (2018)
36. Yu, C., Wang, J., Gao, C., Yu, G., Shen, C., Sang, N.: Context Prior for Scene Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12416–12425 (2020)
37. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
38. Nguyen, T.M., Wu, Q.J.: Fast and robust spatially constrained Gaussian mixture model for image segmentation. IEEE Trans. Circuits Syst. Video Technol. **23**(4), 621–635 (2013)
39. Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., Garcia-Rodriguez, J.: A survey on deep learning techniques for image and video semantic segmentation. Appl. Soft Comput. **70**, 41–65 (2018)
40. Fritsch, J., Kuehnl, T., Geiger, A.: A new performance measure and evaluation benchmark for road detection algorithms. In: 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), pp. 1693–1700. IEEE (2013)

Jingyi Liu received the B.S. degree in mechanical engineering and automation and the M.S. degree in mechanical and electronic engineering from Shanghai University, Shanghai, China, in 2014 and 2017, respectively, where he is currently working toward the Ph.D. degree. His research interests include computer vision, machine learning, and perception for mobile robotics.

Hengyu Li received the B.S. degree in mechanical engineering and automation from Henan Polytechnic University, China, in 2006, and the M.S. and Ph.D. degrees in mechanical and electronic engineering from Shanghai University, China, in 2009 and 2012, respectively, where he is currently an Associate Professor with the School of Mechatronic Engineering and Automation. His research interests include mechatronics and robot bionic vision systems and autonomous cooperative control for multiple robots.

Jun Luo is the Director of State Key Laboratory of Mechanical Transmission with Chongqing University, and he is a Professor in the School of Mechatronic Engineering and Automation of Shanghai University, China. He received the B.S. degree in Mechanical Engineering from Henan Polytechnic University, Jiaozuo, China, in 1994, the M.S. degree in Mechanical Engineering from Henan Polytechnic University, Jiaozuo, China, in 1997, and the Ph.D. degree in Research Institute of Robotics in Shanghai Jiao Tong University, Shanghai, China, in 2000. His research areas include robot sensing, sensory feedback, mechatronics, man-machine interfaces, and special robotics.

Shaorong Xie received the B.S. and M. S. degrees in mechanical engineering from Tianjin Polytechnic University, Tianjin, China, in 1995 and 1998, respectively, and the Ph.D. degree in mechanical engineering from the Institute of Intelligent Machines at Tianjin University and the Institute of Robotics and Automatic Information System, Nankai University, Tianjin, China, in 2001. She is a Professor with the School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China. Her research areas include advanced robotics technologies, bionic control mechanisms of eye movements, and image monitoring systems.

Yu Sun received the B.S. degree in electrical engineering from the Dalian University of Technology, Dalian, China, in 1996, the M.S. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1999, and the M.S. degree in electrical engineering and the Ph.D. degree in mechanical engineering from the University of Minnesota, Minneapolis, MN, USA, in 2001 and 2003, respectively. He is currently a Professor with the University of Toronto, Toronto, ON, Canada. He is a Tier I Canada Research Chair and the Director of the Robotics Institute. His lab specializes in developing innovative technologies and instruments for manipulating and characterizing cells, molecules, and nanomaterials. Prof. Sun was elected Fellow of American Society of Mechanical Engineers (ASME), American Association for the Advancement of Science (AAAS), National Academy of Inventors (NAI), Canadian Academy of Engineering (CAE), and Royal Society of Canada (RSC) for his work on micro-nano devices and robotic systems.

Affiliations

Jingyi Liu[1] · Hengyu Li[1] · Jun Luo[1] · Shaorong Xie[1] · Yu Sun[2]

Jingyi Liu
luckyi@shu.edu.cn

Jun Luo
luojun@shu.edu.cn

Shaorong Xie
srxie@shu.edu.cn

Yu Sun
sun@mie.utoronto.ca

[1] School of Mechatronic Engineering and Automation, Shanghai University, 99 Shangda Road, Shanghai, 200444, China

[2] Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Ontario, Canada