A Deep Network for Object Detection on Inland Waters

Supplementary Material

Dennis Griesser[1,2] Bastian Goldluecke[1] Matthias O. Franz[2] Georg Umlauf[2]
[1]University of Konstanz [2]HTWG Konstanz
{dennis.2.griesser, bastian.goldluecke}@uni-konstanz.de
{mfranz, umlauf}@htwg-konstanz.de

1. Stereo object detection baseline

As a baseline method, we adopt a classical maritime stereo object detection approach [4] that estimates the water surface from the reconstructed point cloud and subsequently detects objects based on the points above this surface. In the following, we describe the modifications required to adapt the method to the Lake Constance Obstacle Detection dataset [3] and provide details on the implementation.

The baseline was implemented by using the water surface detector introduced in [2]. Subsequently, the point cloud is reconstructed using semi-global block matching. Only points that lie at least 25cm above the detected water surface and do not exceed a height of 3m are retained. In contrast to [4], we define only a two-dimensional grid along the x-axis from -100m to 100m and the z-axis from 0m to 250m, with a cell resolution of 0.5m. For each grid cell, the corresponding point count is stored. Then a morphological closing operation with a structuring element of size 7 is applied to close small gaps. To reduce noise, a mean filter with a kernel size of 3 is used. The binary occupancy map is created by applying a threshold of 5 points per cell. Connected components are identified from this binary map, and for each connected region, principal component analysis (PCA) is performed on its convex hull to derive the final bounding boxes. To enable the computation of metrics such as average precision, a detection score is defined as a weighted sum of a shape, a distance, and a point score

$$0.4 \cdot s_{\text{shape}} + 0.2 \cdot s_{\text{dist}} + 0.4 \cdot s_{\text{points}}.$$

The first two scores are given by the eigenvalues μ_1, μ_2 of the PCA and the center c of the bounding box

$$s_{\text{shape}} = 1 - \mu_2/(\mu_1 + \mu_2), \quad s_{\text{dist}} = 1/(1 + \|\mathbf{c}\|),$$

while s_{points} is computed as the number of points within the bounding box divided by the constant 300.

2. Water surface estimation

During the test phase or in practical applications, a module is required to estimate the water surface. For this purpose, we use an algorithm [2] that identifies all pixels belonging to the water surface and fits a plane to the corresponding points in the stereo point cloud to approximate the water surface. This algorithm estimates the rigid body motion $\mathbf{R}^1_{\approx}, \mathbf{t}^1_{\approx}$, allowing us to sample points on the water surface.

3. Hyperparameter selection

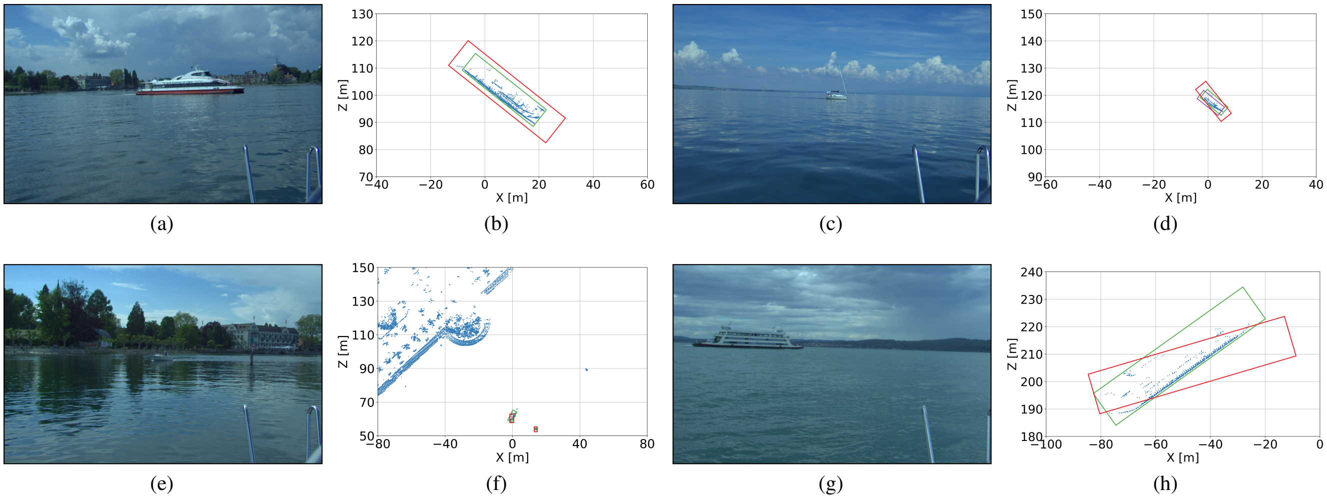
For the ablation studies on the KITTI dataset, we adopted the same 3D space discretization hyperparameters as those used in the DSGN [1] paper. For the 3D space discretization in the inland water experiments, the parameters $x_{min}, x_{max}, z_{min}, z_{max}$ were chosen to cover the entire frustum for which annotations are available. We set $y_{min} = 1$ and $y_{max} = -3$, since points are sampled on the water surface (with the y-axis pointing downward), and to ensure that objects are also captured even if the water surface is not accurately estimated. $\Delta x = 0.25$ was chosen to allow detection of thin objects, such as piles, while $\Delta y = 0.2$ and $\Delta z = 0.5$ were selected to limit memory consumption while still maintaining a reasonable sampling resolution for objects.

4. Qualitative results DSGN baseline (D)

In the paper, qualitative results were presented in comparison to the classical maritime stereo method (B). Here, we additionally visualize in Figure 1 the qualitative results using the DSGN baseline (D).

5. Hardware

The networks were trained on a GPU server equipped with four NVIDIA A100 GPUs (40 GB each).



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 1. Qualitative results of our network (red bounding boxes) compared to the DSGN [1] baseline (D) (violet bounding boxes) and ground truth (green bounding boxes). Images (a), (c), (e), and (g) show the left camera view. The corresponding BEV detections are in (b), (d), (f), and (h), zoomed to the region of interest. Note that the baseline detects only the sailboat under bare poles in c), d).

References

[1] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. DSGN: Deep stereo geometry network for 3d object detection. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2020.

[2] Dennis Griesser, Georg Umlauf, and Matthias O. Franz. Visual pitch and roll estimation for inland water vessels. In *2023 IEEE International Conference on Robotics and Automation*, pages 1961–1967, 2023.

[3] Dennis Griesser, Matthias O. Franz, and Georg Umlauf. Enhancing inland water safety: The lake constance obstacle detection benchmark. In *IEEE International Conference on Robotics and Automation*, pages 14808–14814, 2024.

[4] Jon Muhovič, Rok Mandeljc, Borja Bovcon, Matej Kristan, and Janez Perš. Obstacle tracking for unmanned surface vessels using 3-d point cloud. *IEEE Journal of Oceanic Engineering*, 45(3):786–798, 2020.