The MaSTr1325 dataset for training deep USV obstacle detection models

Borja Bovcon[1], Jon Muhovič[2], Janez Perš[2] and Matej Kristan[1]

***Abstract*— The progress of obstacle detection via semantic segmentation on unmanned surface vehicles (USVs) has been significantly lagging behind the developments in the related field of autonomous cars. The reason is the lack of large curated training datasets from USV domain required for development of data-hungry deep CNNs. This paper addresses this issue by presenting MaSTr1325, a marine semantic segmentation training dataset tailored for development of obstacle detection methods in small-sized coastal USVs. The dataset contains 1325 diverse images captured over a two year span with a real USV, covering a range of realistic conditions encountered in a coastal surveillance task. The images are per-pixel semantically labeled. The dataset exceeds previous attempts in this domain in size, scene complexity and domain realism. In addition, a dataset augmentation protocol is proposed to address slight appearance differences of the images in the training set and those in deployment. The accompanying experimental evaluation provides a detailed analysis of popular deep architectures, annotation accuracy and influence of the training set size. MaSTr1325 will be released to reaserch community to facilitate progress in obstacle detection for USVs.**

***Index Terms*— USVs, obstacle detection, dataset, CNNs**

I. Introduction

Small-sized unmanned surface vehicles (USVs) are affordable devices for automated inspection of hazardous areas and periodic surveillance of coastal waters. A high level of autonomy with a considerable focus on timely detection and avoidance of nearby obstacles is required for practical use. A particularly appealing and low-cost obstacle detection mechanism is semantic interpretation of images captured by on-board cameras [1], [2], [3].

Numerous image processing methods have been proposed for obstacle detection in robotics. In particular, in the related field of unmanned ground vehicles (UGVs), approaches based on semantic segmentation [4], [5], [6], [7], [8] have shown excellent performance over the last few years. The progress has been primarily driven by the design and availability of large publicly available training sets like BDD100k [9], KITTI [10] and Cityscapes [11], which provided the required labeled examples for the notoriously data-hungry deep segmentation architectures. These approaches demonstrated a substantial capacity to model various semantic textures important for UGV scene understanding.

*This work was supported in part by the Slovenian research agency (ARRS) programmes P2-0214 and P2-0095, and the Slovenian research agency (ARRS) research project J2-8175.

[1]Borja Bovcon and Matej Kristan are with University of Ljubljana, Faculty of Computer and Information Science, Slovenia
{borja.bovcon, matej.kristan}@fri.uni-lj.si

[2]Jon Muhovič and Janez Perš are with University of Ljubljana, Faculty of Electrical Engineering, Slovenia
{jon.muhovic, janez.pers}@fe.uni-lj.si

Fig. 1 *A USV was used to acquire the MaSTr1325 dataset [1] in the coastal waters of Koper, Slovenia (upper row). Highly detailed pixel labels for water (cyan), sky (deep blue) and obstacles (yellow) are provided (bottom row). Pixels at the seams between regions are labelled as uncertain (turquoise).*

The state of obstacle detection by semantic segmentation in USVs lags far behind that on the UGV domain. Only few attempts have been made to evaluate popular CNN architectures for obstacle detection on maritime domain ([12], [13]). A major drawback of the existing attempts is the lack of curated domain-specific large training sets that would allow training these networks and to study them on the USV domain. The latter introduces segmentation challenges which are different from the UGV domain. For instance, navigable surface of USVs (water) is visually very inconsistent due to its dynamic appearance, influence of the weather and reflection compared to navigable surface of UGVs (road). Furthermore, obstacles in the marine environment have a more diverse appearance than on the road, and in addition, even small, submerged obstacles might present a significant threat to the USV. Thus state-of-the-art architectures from UGVs might not be appropriate for use on USVs.

This paper addresses the lack of curated training datasets for deep segmentation methods on the USV domain. Our major contribution is a new diverse training dataset captured by our USV (see Figure 1 top left) under variety of weather conditions expected in a typical coastal surveillance task. The dataset, MaSTr1325, contains 1325 manually per-pixel annotated images with labels important for the task of image-based obstacle detection. MaSTr1325 is currently by far the largest and most detailed dataset of its kind for USV obstacle

[1] The MaSTr1325 dataset is made publicly available at: https://vicos.si/Projects/Viamaro

detection. We consider a realistic situation that in practice color grading and camera exposure at application time might be different from that in the training set. We propose a basic color transfer data augmentation protocol to address this. Three popular deep learning semantic segmentation architectures are trained on 53000 images of the augmented MaSTr1325 and evaluated on the task of obstacle detection. A detailed analysis of architectural choices is provided and the impact of the dataset size, annotation accuracy and the proposed color transfer augmentation is experimentally evaluated. The dataset, augmentation protocol and evaluation protocol will be publicly released to the research community.

II. Related work

It has been a common practice to use range sensors, such as RADAR [14], [15], SONAR [16] or LIDAR, for obstacle detection in a marine environment. However, range sensors are typically large in size and are unable to discriminate between water and land in the far field [17]. Recently, cameras, combined with computer vision algorithms, are gaining prominence as an affordable, lightweight [2] and very powerful obstacle detection mechanism.

Numerous image-processing methods for obstacle detection have been proposed. Prasad *et al.* [3] have evaluated the performance of various state-of-the-art background subtraction methods as obstacle detection mechanisms on a marine domain. In-depth experiments on Singapore Marine Dataset (SMD) [18] have shown that deceptive dynamics of water causes a substantial amount of false positive detections. Wang *et al.* [19] introduced a 3-D reconstruction based method for obstacle detection. However, 3-D reconstruction methods are only capable of detecting obstacles that significantly protrude through the water surface and not partially submerged ones. Moreover, in the state of a calm sea, where water lacks a texture, a 3-D reconstruction is severely degraded, leading to false detections. On the other hand, Kristan *et al.* [1] and Bovcon *et al.* [20] proposed a graphical model for obstacle detection via semantic segmentation. Although their method is capable of detecting obstacles protruding through the water surface as well as the floating ones, it still fails in the presence of artifacts on water (small wakes, etc.)

Inspired by remarkable results of deep learning methods, developed for the UGV domain, Lee *et al.* [21] proposed using a general Faster R-CNN [22] to detect and classify seven different types of ships. However, their method is unable to detect arbitrary obstacles in the water without providing a huge amount of additional training data. Jeong *et al.* [23] have used PSPNet [24] (pretrained on ADE20k dataset) to segment the image and use the extracted water component for horizon approximation. Recently, Cane *et al.* [12] evaluated three deep semantic segmentation networks (SegNet [25], ENet [26], ESPNet [27]) for object detection in maritime surveillance. Similarly to [23], they have trained on ADE20k dataset. The original 150 classes of ADE20k were mapped to one of four custom classes (sea, sky, object and other). Despite the fact that ADE20k dataset has well over 20000 images in total, only a small subset of 448 images contains visible sea. Moreover, some of these images are aerial shots or water close-ups, which do not portray a realistic viewpoint of USVs and are detrimental to the learning of the deep segmentation method.

Fig. 2 *MaSTr1325 captures a variety of weather conditions ranging from foggy, partly cloudy with sunrise, overcast to sunny (top) and visually diverse obstacles (bottom).*

The performance of the deep learning methods is strongly correlated with the quality and quantity of learning cases. Although several publicly available maritime datasets exist (Modd [1], Modd2 [28], IPATCH [29], SEAGULL [30], SMD [18]) none of them has pixel-wise ground truth annotations, required for learning. A recent attempt to construct a marine environment training set was made in [13]. However, the dataset contained less than 300 images, which is too small for effective training of the modern deep learning models.

III. Marine segmentation dataset – MaSTr1325

To reflect the realism of the USV environment encountered on an average mission, the dataset was captured by a real USV over a span of 24 months in the gulf of Koper, Slovenia. A small-sized $2\,\mathrm{m}$ long USV, developed by Harpha Sea, d.o.o., was used (Figure 1 top left). The USV uses a steerable thrust propeller for guidance, and can reach the maximum velocity of $2.5\,\mathrm{m\,s^{-1}}$. It is equipped with LIDAR, compass, GPS, IMU unit, two side cameras and a main stereo camera system *Vrmagic VRmMFC*, which comprise of two *Vrmagic VRmS-14/C-COB* CCD sensors with baseline $0.3\,\mathrm{m}$, *Thorlabs MVL4WA* lens with $3.5\,\mathrm{mm}$ focal length, maximum aperture of $\mathrm{f}/1.4$, and a $132.1°$ FOV.

The stereo system is mounted approximately $0.7\,\mathrm{m}$ above the water surface, faces forward and is z-axis aligned with the IMU. Cameras are connected to the on-board computer through USB-2.0 bus which restricts the data flow to 10 frames per second at resolution 1278×958 pixels. The aperture of cameras is automatically adjusted according to the lighting conditions, preventing underexposure and occurrence of indistinguishable dark areas in shades.

From out of approximately fifty hours of footage, captured in the two years, we have hand-picked representative images of a marine environment. A particular attention was paid to include various weather conditions and times of day to ensure the variety of the captured dataset. Finally, 1325 images were accepted for the final dataset. Examples are shown in Figure 2.

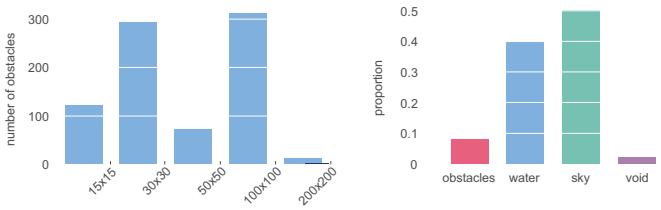

Fig. 3 *Size distribution of obstacles in the training set, based on obstacle area measured in pixels (left) and average proportions of annotated pixels for each category (right).*

Each image from the dataset was manually annotated by human annotators with three categories (sea, sky and environment). An image editing software supporting multiple layers per image was used and each semantic region was annotated in a separate layer by multiple sized brushes for speed and accuracy. All annotations were carried out by in-house annotators and were verified and corrected by an expert to ensure a high-grade per-pixel annotation quality. The annotation procedure and the quality control took approximately twenty minutes per image. To account for the annotation uncertainty at the edge of semantically different regions, the edges between these regions were labeled by the "unknown" category (see Figure 1 bottom). This label ensures that these pixels are excluded from learning.

Figure 3 shows a distribution of obstacle sizes and proportion of annotated pixels for each category. The obstacles cover a wide range of sizes. The majority of pixels are annotated by water or sky, which comes from the nature of the task. Still, a considerable number of obstacle pixels are present in the dataset, covering a large variety of obstacle appearances (see Figure 2 for examples).

A. Augmentation and adaptation to target domain

In practice, we can realistically expect that the camera exposure and other parameters at application time might differ from those used in the dataset acquisition. In addition, for training deep networks, a useful approach is to augment the dataset by generating geometric perturbations of the original images, effectively inducing a regularization effect in the learning.

The following augmentation protocol is thus introduced that considers the target sensors in addition to standard augmentation. We propose applying vertical mirroring and central rotation of $\pm\{5, 15\}$ degrees. In addition, the color augmentation is applied to each image, where a small set of descriptive images from target domain is used for color transfer following the approach [31]. An example of dataset augmentation is shown in Figure 4.

IV. Performance evaluation protocol

The Modd2 [28] is chosen for performance evaluation of the networks trained on MaSTr1325, since it is currently the largest and the most challenging USV obstacle detection dataset. It consists of 11675 rectified stereo images captured

Fig. 4 *An image from MaSTr1325 (top left), followed by its seven color augmentations and four rotation augmentations in the last row.*

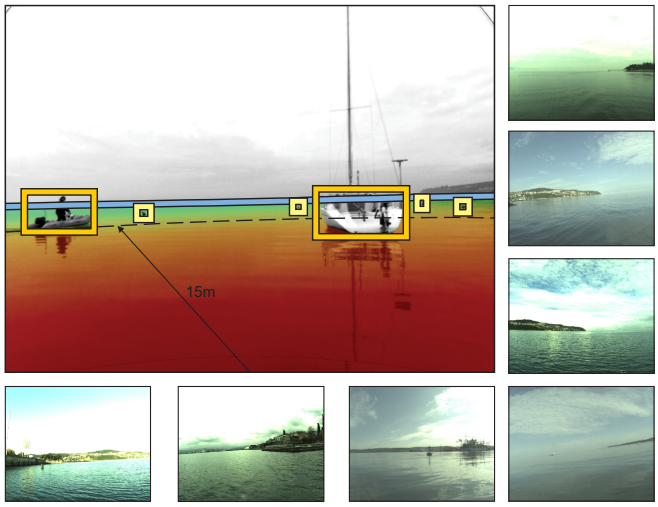

Fig. 5 *Objects below water edge are annotated by bounding boxes, while the edge of the water is annotated separately. We add the 15m danger zone visualized by a color gradient, ranging from red (dangerous) to green (safe) based on distance from the USV. Surrounding smaller images were used for color transfer augmentation.*

in the coastal waters by the USV described in Section III. Obstacles and water-edge in Modd2 were manually annotated with bounding boxes and a polygon respectively. Timely and accurate obstacle detection is of central importance for autonomous navigation. For faster inference, the Modd2 images are resized to the resolution 512×384 pixels.

The accuracy of semantic segmentation is reported by three metrics inspired by Long *et al.* [32]:

1) Mean pixel accuracy: $\frac{\sum_i n_{ii}}{\sum_i t_i}$
2) Mean IOU: $\frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$
3) Frequency weighted IOU: $\left(\sum_k t_k\right)^{-1} \sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$

where n_{cl} denotes the number of classes in the ground truth, n_{ji} represents the number of pixels of the class j predicted to belong to the class i, while t_i stands for the total number of pixels of class i in the ground truth segmentation. These

segmentation metrics, however, lack insight into water-edge and obstacle detection accuracy. For this task we use the obstacle-detection metrics proposed by [28], where water-edge accuracy is measured by mean-squared error over all sequences, while the accuracy of detected obstacles is measured by the number of true positives (TP), false positives (FP), false negatives (FN) and by the overall F-measure, i.e., a harmonic mean of precision and recall.

Not all obstacles in the scene present equal danger to the USV. For example, obstacles in close proximity are more hazardous than distant ones. To address this, we propose a danger-zone mask for sequences from Modd2. Assuming an average speed of $1.5\,\mathrm{m\,s^{-1}}$, we define the danger zone as the farthest point reachable within $10\,\mathrm{s}$. Thus the danger-zone is defined as a radial area, centered at the location of USV with radius $15\,\mathrm{m}$ (shown in Figure 5). We calculated the danger-zone for each image in Modd2 from the IMU.

V. Experimental evaluation

A summary of the tested baseline segmentation CNNs and training implementation details are given in Sections V-A and V-B. Section V-C benchmarks these architectures and selects the most promising for the USV domain. The architectural elements are further empirically analyzed in Section V-D. The influence of the training set size and annotation properties on the USV domain are analyzed in Sections V-E and V-F.

A. Baseline CNN segmentation architectures

Three popular CNN architectures, that have shown excellent performance in the segmentation domain and form the basis of many UGV state-of-the-art segmentation CNNs have been selected as baseline segmentation methods in our benchmark:

1) U-Net [33] was initially designed for bio-medical image segmentation, but it achieves state-of-the-art results across a wider range of domains [34], [35], [36] as well. Its architecture contains an encoder which captures context and a symmetric decoder that provides precise localization. Repeated convolution and max pooling layers gradually reduce the feature channels' width and height and increase their number. The decoder then gradually up-samples the channels by additional convolutions followed by bilinear interpolation. The channel width/height is thus doubled at each step and the number of channels is halved. Skip connections with concatenation are used to maintain details and prevent information loss due to size reduction.

2) PSPNet [24] was designed as a general scene parsing network. It uses a pre-trained ResNet-50 [37] backbone with atrous convolutions to increase the filters receptive field size without increasing the number of parameters. Feature maps are fed to the pyramid-pooling module. The module consists of four pyramid levels/scales. After each pyramid level, an additional convolution is applied to reduce the dimension of feature map. Bilinear interpolation is used to up-sample low dimensional feature maps. Up-sampled feature maps from the pyramid-pooling-module are concatenated with the initial feature map and sent through a convolution to generate the final prediction map, which is further up-sampled to the original image size.

3) DeepLabv2 [38] uses a ResNet-101 [37] backbone with atrous convolutions and Atrous Spatial Pyramid Pooling (ASPP) to extract features. ASPP enables to capture objects and context at different scales by using various sampling rates to enlarge the receptive field. The result is fed to a fully connected Conditional Random Field (CRF), which is separately tuned, to improve the object localization. A multi-scale version of DeepLabv2 passes multiple re-scaled versions of the input image to parallel branches, rescales responses before CRF and adds them. In the following, the single-scale and multiple-scale versions of the DeepLabv2 are denoted by DLv2s and DLv2m, respectively.

B. Implementation details

A softmax cross-entropy loss with momentum 0.9 were used to train the segmentation networks. The initial learning rate was set to 10^{-4} and the standard polynomial reduction decay 0.9 was applied. The backbones in PSPNet and DeepLabv2 were intialized by ResNet-50 and ResNet-101 pretrained on ImageNet [39]. Fine-tuning on our dataset was carried out for 13 epochs.

The MaSTr1325 dataset was first augmented by applying central rotations, generating in total 5300 new images with accurate ground truth. The second part of augmentation was done by color properties transfer. Seven representative images were selected from Modd2 [28] that depict the variability of weather and lighting conditions present in test sequences (shown in Figure 5). These images were used for color properties transfer approach (Section III), generating 7 new images for each image in the already augmented dataset. Thus the overall size of training images from MaSTr1325 along with augmentation amounted to 53000 images.

The tested methods were implemented in Tensorflow [40] and run on a desktop computer with Intel Core i7-7700 $3.6\,\mathrm{GHz}$ CPU and nVidia GTX1080 Ti GPU.

C. Baseline CNN segmentation architecture benchmark

All four baseline architectures, U-Net [33], PSP-Net [24], DLv2s and its multi-scale counterpart DLv2m, were trained on augmented MaSTr1325 and evaluated on the *left image* Modd2 [28] stereo pairs. According to traditional semantic segmentation measures reported in Table I, a single-scale DLv2 outperforms all other architectures, including a multi-scale version. PSPNet, which is considered a state-of-the-art in UGV segmentation achieves worst results. Ranking consistency was tested by evaluating the methods on the *right images* of Modd2. The performance of all methods is consistent in both views, thus the ranking is stable.

Additional insights are shed by the Modd2 obstacle detection measures in Table II. The water edge is most accurately estimated by DLv2 variants, followed by U-Net and PSP, whose error is approximately twice as large (Figure 6 bottom row). U-Net obtains the highest number of true positives, followed by DLv2 variants and PSPNet. A visual inspection

TABLE I *Semantic segmentation metrics. All results are reported in percentages.*

	architecture	Mean PA	Mean IOU	fw-IOU
Left camera	U-Net [33]	91.96	96.80	90.65
	PSPNet [24]	90.98	94.50	88.22
	DeepLabv2s [38]	**92.65**	**97.49**	**91.48**
	DeepLabv2m [38]	92.61	97.41	91.38
Right camera	U-Net [33]	91.91	96.72	90.58
	PSPNet [24]	91.00	94.51	88.25
	DeepLabv2s [38]	**92.58**	**97.39**	**91.35**
	DeepLabv2m [38]	**92.58**	97.36	91.33

TABLE II *Water-edge estimation error μ_{edg}, the number of true positive (TP), false positive (FP), false negative (FN) detections and the F-measure and Time in FPS are computed on entire image (upper part) and separately within the danger zone (bottom part).*

architecture	μ_{edg}	TP	FP	FN	F-measure	Time
U-Net [33]	18.8	**592**	3706	**87**	23.8	15.37
PSPNet [24]	40.0	346	**54**	333	**64.1**	**17.20**
DLv2s [38]	**16.6**	369	108	310	63.8	1.62
DLv2m [38]	**16.6**	304	65	375	58.2	1.39
U-Net [33]	/	**208**	3337	**43**	11.0	15.37
PSPNet [24]	/	126	39	125	60.6	**17.20**
DLv2s [38]	/	167	36	84	**73.6**	1.62
DLv2m [38]	/	147	**14**	104	71.4	1.39

reveals that U-Net is highly sensitive to the input and significantly over-fragments the water region (Figure 6 middle row), resulting in the highest number of false positives and the lowest F-measure. The best overall F-measure is achieved by PSPNet, closely followed by DLv2. PSPNet also by far surpasses DLv2 in speed, which is important for practical applications in real-time systems like USV. However, within the danger zone, the DLv2 achieves the highest number of true positives, the lowest number of false detections and significantly outperforms the PSPNet in F-measure. DLv2m, obtains a smaller number of false positives than the single-scale version, but at a cost of reduced true positives, leading to a smaller F-measure. A closer visual inspection of segmentation results showed that the multi-scale version loses small obstacles (Figure 6 top row), which is critical for practical early obstacle detection within the danger zone.

From these results, we selected a single-scale DLv2 architecture for its good trade-off between reliable obstacle detection and low false alarm rate and accurate water edge estimation for further architectural analysis and potential speedups.

D. Ablation study of the DeepLabv2 architecture

The single scale DLv2 architecture is composed of several elements that contribute to computational complexity and performance: (i) the back-bone ResNet-101, (ii) atrouse convolutions in the backbone, (iii) atrouse spatial pyramid pooling (ASPP) and (iv) a fully connected conditional random field (CRF) on the output. The experiment from Section V-C was repeated for variations of the original network to evaluate contribution of each architectural element: a version without CRF ($\text{DL}_{\setminus\text{CRF}}$), a version without ASPP ($\text{DL}_{\setminus\text{ASPP}}$), a version without atrouse convolutions ($\text{DL}_{\setminus\text{ATR}}$) and a version with the backbone replaced by ResNet-50 (containing dilated convolutions) (DL_{Res50}).

TABLE III *Water-edge estimation error μ_{edg}, the number of true positive (TP), false positive (FP), false negative (FN) detections, F-measure and Time in FPS. Detections in danger zone are written in brackets.*

architecture	μ_{edg}	TP	FP	FN	F-measure	Time
$\text{DL}_{\setminus\text{CRF}}$	**16.0**	**436** [**187**]	98 [80]	**243** [**64**]	**71.9** [**72.2**]	**17.21**
$\text{DL}_{\setminus\text{ASPP}}$	17.8	339 [129]	115 [53]	340 [122]	59.8 [59.6]	1.51
$\text{DL}_{\setminus\text{ATR}}$	24.6	250 [99]	**82** [**19**]	429 [152]	49.5 [53.7]	1.50
DL_{Res50}	18.6	335 [154]	183 [78]	344 (97)	56.0 [63.8]	1.54

TABLE IV *Segmentation results without data augmentation with respect to the number of training images N_{train}, measured by mean PA, mean IOU and fw-IOU.*

N_{train}	Mean PA	Mean IOU	fw-IOU
200	92.41	97.04	91.07
500	92.46	97.23	91.20
800	92.51	97.33	91.30
1000	92.51	97.36	91.31
1325	92.52	97.38	91.33
1325_{Aug}	**92.65**	**97.49**	**91.48**

Results in Table III show that all elements of DLv2 contribute to performance, except the CRF. In fact, compared to the orignal DLv2 performance (Table II) the performance is boosted across all measures when removing the CRF. $\text{DL}_{\setminus\text{CRF}}$ achieves improved overall F-measure at a significant speed improvement, surpassing that of the PSPNet. In addition to atrouse convolutions and ASPP, the backbone seems to play a crucial role. A significant performance drop is observed when replacing the ResNet-101 with a shallower equivalent, meaning that sufficient depth is required to sufficiently well encode the various local water appearances.

E. Influence of the training set size and data augmentation

Deep learning architectures are notoriously data hungry and their performance is importantly affected by the training data set size. To evaluate the effectiveness of the proposed training set size and of the data augmentation protocol from Section III, the DeepLabv2 without CRF ($\text{DL}_{\setminus\text{CRF}}$) was re-evaluated with increasing training set size. The results are reported in Tables IV and V.

The overall segmentation accuracy (Table IV) steeply increases up to 800 training images and gradually saturates. The same trend is observed in water edge estimation accuracy (Table V), since the latter is directly related with the overall segmentation.

A somewhat different trend is observed in detection-based measures (Table V). The percentage of true positives appears

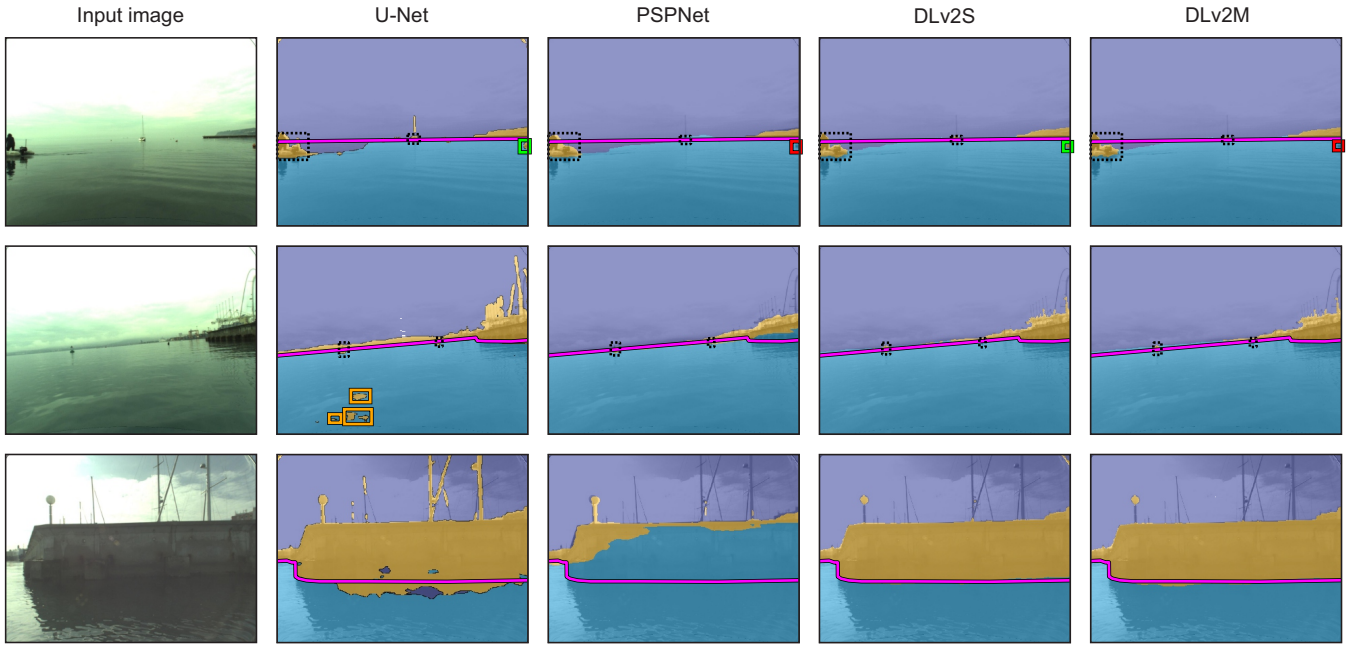

Fig. 6 *Qualitative comparison of the baseline architectures. The sky, obstacles and water components are denoted with deep-blue, yellow and cyan color, respectively. The ground truth sea edge is annotated with a pink line, while ground truth obstacles are outlined with a dotted bounding box. False positives are marked with an orange bounding box, false negatives are marked with a red bounding box, whereas correctly detected obstacles are marked with a green bounding box.*

TABLE V *Water-edge estimation error μ_{edg}, the number of true positive (TP), false positive (FP), false negative (FN) detections and F-measure. Detections in danger zone are written in brackets.*

N_{train}	μ_{edg}	TP	FP	FN	F-measure
200	17.5	362 [146]	**45** [**29**]	317 [105]	66.7 [68.5]
500	17.1	410 [178]	70 [43]	269 [73]	70.8 [**75.4**]
800	16.7	392 [171]	87 [66]	287 [80]	67.7 [70.1]
1000	16.7	390 [174]	102 [76]	289 [77]	66.6 [69.5]
1325	16.5	409 [**188**]	145 [106]	270 [**63**]	66.3 [69.0]
1325_{Aug}	**16.0**	**436** [187]	98 [80]	**243** [64]	**71.9** [72.2]

to be constantly increasing with the dataset size. However, so does the number of false positives. This might be counter intuitive at first glance, but a careful visual inspection of the false positives in the images revealed that these result from mirroring reflections and in some cases parts of sea foam, which are in fact anomalies in the water texture. Examples are shown in Figure 7.

The network thus did improve in learning of the water appearance with increasing the number of training images and was gradually better in separating small obstacles from the water, thus increasing the true positive rate. However, this also made it better in "detecting" parts of mirroring reflections as potential obstacles, which translate into false positives.

The last line in Tables IV and V shows results when using all training images along with the data augmentation protocol from Section III. The overall segmentation accuracy is increased across all measures as well as water edge estimation. More importantly, the augmentation gives a healthy boost in object detection rates by increasing the number of true positives and reducing false positives (last two lines in Table V). The results support the importance of our color-transfer-based data augmentation.

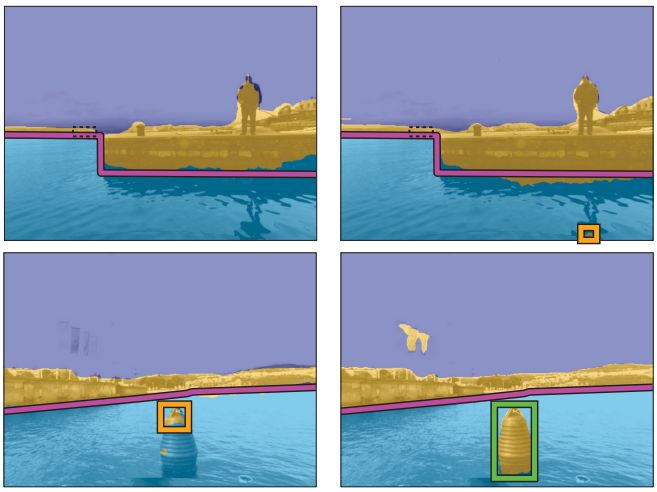

Fig. 7 *A $DL_{\setminus CRF}$ network trained on 200 images (left) is not affected by object mirroring in the water (upper left), but also nearly completely misses a large obstacle (bottom left). The same network trained on 1325 images (right) perfectly detects the obstacle, but also detects a false positive on the object mirrored reflection (upper right).*

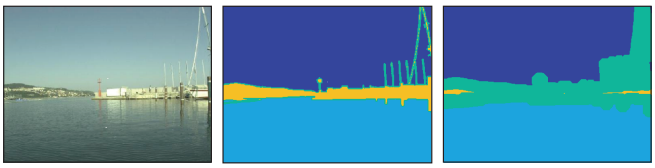

Fig. 8 *Original image (left), detailed pixel-wise annotation (middle) and an approximate annotation (right).*

F. Importance of per-pixel annotation accuracy

Per-pixel semantic segmentation annotation of datasets is an arduous, time-consuming task. Most of the time is consumed in careful annotations of the boundaries between semantically different regions. The annotation time would be drastically decreased if accurate labeling was only required within the semantic region, but could allow annotating a wider area containing the edges of neighboring semantic regions by an "unknown" label – i.e., a label that excludes pixels from learning by setting the error gradients in the back-prop to zero.

To test the feasibility of such approximate annotation, the individual labels in the MaSTr1325 images were dilated by a 11×11 kernel. The dilated pixels that intersected with a different labels were denoted as "unknown" label. An example of thus modified training labels is shown in Figure 8.

A $\text{DL}_{\setminus\text{CRF}}$ network was trained on the modified labels (with augmentation). The network obtained 90.70% in fw-IOU, 19.8px error in water edge accuracy, 198 TP, 68 FP and 41.9% F-measure. Compared to $\text{DL}_{\setminus\text{CRF}}$ trained on original labels (Tables III and IV) a large performance drop is observed. The number of false positives drops as well, but this is attributed to a poorer overall detection rate. Thus the detailed annotation is not only crucial for water edge segmentation, but also for the overall detection of obstacle presence.

VI. Conclusion

We presented MaSTr1325, a comprehensive dataset carefully designed for training deep semantic segmentation methods for obstacle detection in small coastal-line autonomous surface vehicles. This is the largest and most detailed per-pixel labeled dataset of its kind for the coastal water USV obstacle detection task.

The dataset was used to benchmark three popular semantic segmentation architectures. While the different architectures perform comparably in terms of general semantic segmentation measures, a single-scale DeepLabv2 [38] excelled in obstacle-detection measures. Results show that water segmentation proves highly challenging due to its varying appearance. Deep backbone architectures with a high capacity are required to cope with this variation. This was verified in observing performance drops in best-performing architecture when reducing its backbone architecture (Section V-D). A notable result is that conditional random fields often used for segmentation regularization hamper the detection rate on part of missing small obstacles. In fact, we identified the single-scale DeepLabv2 with CRF removed as the minimal architecture with an excellent tradeoff between the number of true positive and false positive detections, achieving a real-time performance, which is crucial for practical use.

Experiments showed that the dataset size is appropriate for training the networks. The segmentation accuracy and detection rate stabilize from 800 images onward, although evidence shows that improvements are expected from future increases of the training set. A significant performance boost is observed in applying our color transfer augmentation protocol, that transfers basic appearance properties from the target data-set to the training set (Section III). Experiments also show that accurate boundary annotation between semantic regions is crucial not only for water edge estimation, but also for increasing the obstacle presence detection (Section V-F). Time-consuming detailed annotation thus, unfortunately, cannot be avoided. The MaSTr1325 dataset along with the evaluation protocol and data augmentation routines will be made publicly available in hope to facilitate progress in autonomous boats perception designs, a domain crucially lacking sufficiently large training datasets and benchmarks.

Empirical results indicate that false detections caused by foam fragments and fragments of obstacles mirroring in the water present a challenge for practical obstacle detection. We believe that these cannot be dealt with static images and temporal component must be considered. Design of efficient architectures informed by the results of this paper and coping with reflections will be the focus of our future work.

References

[1] M. Kristan, V. S. Kenk, S. Kovačič, and J. Perš, "Fast image-based obstacle detection from unmanned surface vehicles," *IEEE TCYB*, vol. 46, no. 3, pp. 641–654, 2016.

[2] J. Larson, M. Bruch, R. Halterman, J. Rogers, and R. Webster, "Advances in autonomous obstacle avoidance for unmanned surface vehicles," SPAWAR San Diego, Tech. Rep., 2007.

[3] D. K. Prasad, C. K. Prasath, D. Rajan, L. Rachmawati, E. Rajabally, and C. Quek, "Object detection in a maritime environment: Performance evaluation of background subtraction methods," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–16, 2018.

[4] J. M. Alvarez, Y. LeCun, T. Gevers, and A. M. Lopez, "Semantic road segmentation via multi-scale ensembles of learned features," in *European Conference on Computer Vision*. Springer, 2012, pp. 586–595.

[5] J. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez, "Road scene segmentation from a single image," in *European Conference on Computer Vision*. Springer, 2012, pp. 376–389.

[6] D. Levi, N. Garnett, E. Fetaya, and I. Herzlyia, "Stixelnet: A deep convolutional network for obstacle detection and road segmentation." in *BMVC*, 2015, pp. 109–1.

[7] G. L. Oliveira, W. Burgard, and T. Brox, "Efficient deep models for monocular road segmentation," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 4885–4891.

[8] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2018.

[9] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, 2018.

[10] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[12] T. Cane and J. Ferryman, "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)*. IEEE, 2018, pp. 1–6.

[13] B. Bovcon and K. Matej, "Benchmarking semantic segmentation methods for obstacle detection on a marine environment," in *CVWW 2019*, 2019.

[14] C. Almeida, T. Franco, H. Ferreira, A. Martins, R. Santos, J. M. Almeida, J. Carvalho, and E. Silva, "Radar based collision detection developments on USV ROAZ II," in *OCEANS - EU*, May 2009, pp. 1–6.

[15] C. Onunka and G. Bright, "Autonomous marine craft navigation: On the study of radar obstacle detection," in *ICCAR 2010*, Dec 2010, pp. 567–572.

[16] H. K. Heidarsson and G. S. Sukhatme, "Obstacle detection and avoidance for an autonomous surface vehicle using a profiling sonar," in *ICRA 2011*, May 2011, pp. 731–736.

[17] L. Elkins, D. Sellers, and W. R. Monach, "The autonomous maritime navigation (AMN) project: Field tests, autonomous and cooperative behaviors, data fusion, sensors, and vehicles," *Journal of Field Robotics*, vol. 27, no. 6, pp. 790–818, 2010.

[18] D. K. Prasad, D. Rajan, L. Rachmawati, E. Rajabally, and C. Quek, "Video processing from electro-optical sensors for object detection and tracking in a maritime environment: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 8, pp. 1993–2016, 2017.

[19] H. Wang and Z. Wei, "Stereovision based obstacle detection system for unmanned surface vehicle," in *ROBIO*, 2013, pp. 917–921.

[20] B. Bovcon and M. Kristan, "Obstacle detection for usvs by joint stereo-view semantic segmentation," 2018.

[21] S.-J. Lee, M.-I. Roh, H.-W. Lee, J.-S. Ha, I.-G. Woo, *et al.*, "Image-based ship detection and classification for unmanned surface vehicle using real-time object detection neural networks," in *The 28th International Ocean and Polar Engineering Conference*. International Society of Offshore and Polar Engineers, 2018.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[23] C. Y. Jeong, H. S. Yang, and K. D. Moon, "Horizon detection in maritime images using scene parsing network," *Electronics Letters*, vol. 54, no. 12, pp. 760–762, 2018.

[24] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[25] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[26] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[27] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 552–568.

[28] B. Bovcon, J. Perš, M. Kristan, *et al.*, "Stereo obstacle detection for unmanned surface vehicles by IMU-assisted semantic segmentation," *Robotics and Autonomous Systems*, vol. 104, pp. 1–13, 2018.

[29] L. Patino, T. Nawaz, T. Cane, and J. Ferryman, "Pets 2017: Dataset and challenge," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 2126–2132.

[30] R. Ribeiro, G. Cruz, J. Matos, and A. Bernardino, "A dataset for airborne maritime surveillance environments," *IEEE Trans. Circuits Syst. Video Technol*, 2017.

[31] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[34] Z. Guo, H. Shengoku, G. Wu, Q. Chen, W. Yuan, X. Shi, X. Shao, Y. Xu, and R. Shibasaki, "Semantic segmentation for urban planning maps based on U-Net," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 6187–6190.

[35] L. Liu and Y. Zhou, "A closer look at U-Net for road detection," in *ICDIP 2018*, vol. 10806. International Society for Optics and Photonics, 2018, p. 108061I.

[36] W. Xia, Z. Chen, Y. Zhang, and J. Liu, "An approach for road material identification by dual-stage convolutional networks," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, July 2018, pp. 7153–7156.

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

[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE TPAMI*, vol. 40, no. 4, pp. 834–848, 2018.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[40] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/