Temporal Context for Robust Maritime Obstacle Detection

Lojze Žust[1] and Matej Kristan[1]

***Abstract*— Robust maritime obstacle detection is essential for fully autonomous unmanned surface vehicles (USVs). The currently widely adopted segmentation-based obstacle detection methods are prone to misclassification of object reflections and sun glitter as obstacles, producing many false positive detections, effectively rendering the methods impractical for USV navigation. However, water-turbulence-induced temporal appearance changes on object reflections are very distinctive from the appearance dynamics of true objects. We harness this property to design WaSR-T, a novel maritime obstacle detection network, that extracts the temporal context from a sequence of recent frames to reduce ambiguity. By learning the local temporal characteristics of object reflection on the water surface, WaSR-T substantially improves obstacle detection accuracy in the presence of reflections and glitter. Compared with existing single-frame methods, WaSR-T reduces the number of false positive detections by 41% overall and by over 53% within the danger zone of the boat, while preserving a high recall, and achieving new state-of-the-art performance on the challenging MODS maritime obstacle detection benchmark.**

I. Introduction

Advances in maritime robotics over the last two decades have fostered an emergence of unmanned surface vehicles (USVs). These autonomous boats range from small vessels used for automated inspection of dangerous areas and automation of repetitive tasks like bathymetry or environmental control, to massive cargo and transport ships. This next stage of maritime automation holds a potential to transform maritime-related tasks and will likely impact the global economy. The safety of autonomous navigation systems hinges on their environment perception capability, in particular obstacle detection, which is responsible for timely reaction and collision avoidance.

Cameras as low-power and information rich sensors are particularly appealing due to their large success in perception for autonomous cars [1], [2]. However, recent works [3], [4] have shown that methods developed for autonomous cars do not translate well to USVs due to the specifics of the maritime domain. As a result, several approaches that exploit the domain specifics for improved detection accuracy have been recently proposed [5], [6], [7], [8]. Since everything but water can be an obstacle, classical detectors for individual obstacle classes cannot address all obstacle types. State-of-the-art methods [5] instead casts maritime obstacle detection as an anomaly segmentation problem by segmenting the image into the water, sky and obstacle categories.

*This work was supported by the Slovenian research agency program P2-0214 and project J2-2506.

[1]Lojze Žust and Matej Kristan are with University of Ljubljana, Faculty of Computer and Information Science, Slovenia {lojze.zust, matej.kristan}@fri.uni-lj.si

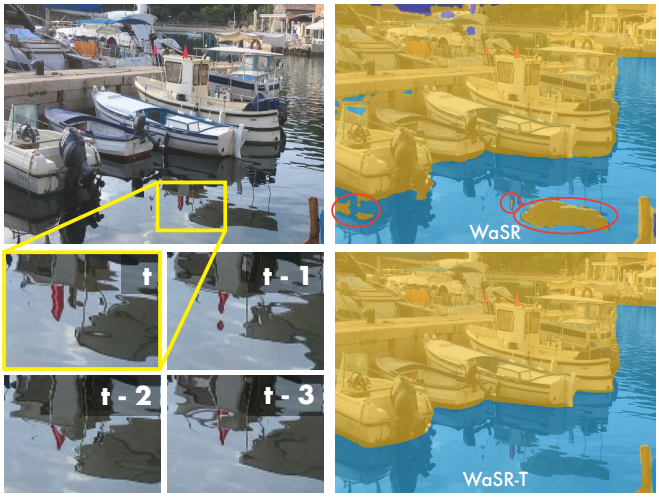

Fig. 1. Single-frame obstacle detection methods (top right) struggle to distinguish between object reflections and true objects. However, reflections exhibit a distinctive temporal pattern compared to true objects (bottom left). WaSR-T (bottom right) considers the temporal context from recent frames to learn these patterns and increase segmentation robustness.

Despite significant advances reported in the recent maritime benchmark [4], the state-of-the-art is still challenged by the reflective properties of the water surface, which cause objects reflections and sun glitter. In fact, given a single image, it is quite difficult to distinguish a reflected object or a spot of sun glitter from a true obstacle (Figure 1). This results in a number of false positive detections, which in practice leads to frequent and unnecessary slowdowns of the boat, rendering current camera-based obstacle detection methods impractical.

We note that while correctly classifying reflections from a single image is challenging, the problem might become simpler when considering the temporal context. As illustrated in Figure 1, due to water dynamics, the reflection appearance is not locally static, like that of an obstacle, but undergoes warped deformations. Based on this insight, we propose a new maritime obstacle segmentation network WaSR-T, which is our main contribution. WaSR-T introduces a new temporal context module that allows the network to extract the temporal context from a sequence of frames to differentiate objects from reflections. To the best of our knowledge, this is the first deep maritime obstacle detection architecture with a temporal component.

We also observe that the challenging maritime mirroring and glitter scenes are underrepresented in the standard training sets. We therefore extend the existing single-frame maritime segmentation training dataset MaSTr1325 [3] with

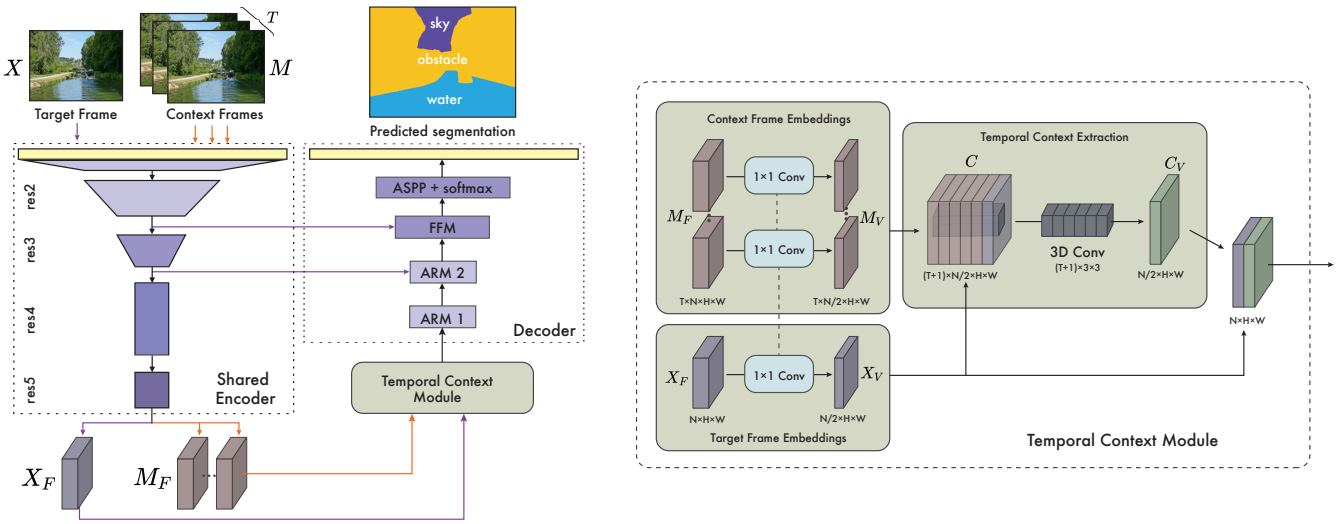

Fig. 2. Overview of WaSR-T (left). Target frame and preceding context frames are fed into a shared encoder producing per-frame feature maps X_F and M_F. The Temporal Context Module (right) extracts the temporal information from per-frame embeddings using a 3D convolution. The resulting temporal context embeddings C_V are combined with target frame embeddings X_V and fed into the decoder which predicts the target frame segmentation.

corresponding preceding frames and introduce additional training images representing challenging reflection conditions, which is our secondary contribution. To maintain the notation convention, we name the extended dataset MaSTr1478. Experiments show that the dataset extension delivers significant performance improvement. Results on the recent maritime benchmark MODS [4] show that, compared to the single-frame WaSR [5], the proposed WaSR-T reduces the number of false positive detections by 30% with a low computational overhead and sets a new state-of-the-art in maritime obstacle detection.

In summary, our main contributions are: (i) WaSR-T, a temporal extension of WaSR [5] that leverages the temporal context for increased robustness and (ii) MaSTr1478, an extension of the existing single-frame training dataset [3] with challenging reflection scenes that facilitates the training of temporal maritime segmentation networks. The new dataset and the WaSR-T source code will be publicly released to facilitate further research of temporal features in maritime obstacle detection.

II. Related work

Semantic segmentation has become a common approach for obstacle detection in the marine domain [3], [6], [4], as it can address both dynamic (*e.g.* boats, swimmers, buoys) and static obstacles (*e.g.* shore, piers, floating fences) in a unified way by posing the problem as anomaly segmentation. Recently, several specialized networks for the marine domain have been proposed for this task [5], [7], [9]. These methods address reflections and increase detection robustness in multiple ways, including regularization techniques [7], specialized loss functions [5] and obstacle-oriented training regimes [8].

However, robustness to reflections is still lacking and causes comparatively low performance within the 15m area near the boat [4], where segmentation errors are most critical. In practice, obstacle detection methods receive frames sequentially, thus the temporal component of the data is also available and could be used to distinguish between reflections and objects. So far, the additional temporal information has not yet been explored in context of maritime obstacle detection.

In other domains with similar access to sequential image data, effort has been made to harness the temporal information to improve the segmentation performance. Some approaches investigate the use of temporal information only during training to improve the temporal consistency of single-frame networks. [10] and [11] achieve this by propagating the segmentation masks in consecutive frames by optical flow.

Incorporating temporal information into the network for improved prediction has been explored as well, with attention-based approaches being the most prevalent method. In video object segmentation [12], [13], [14] attention is used to aggregate the information from features and segmentation masks of previous ”memory” frames based on the attention between the target and memory features. However, these methods are designed mainly for propagating initial segmentation masks of large foreground objects over the video sequence and are not directly suitable for general purpose discriminative semantic segmentation required for obstacle detection.

Similarly, in video semantic segmentation [15], [16] attention-based approaches are used to aggregate the temporal information from recent frames to improve general purpose semantic segmentation. [16] additionally introduces auxiliary losses, which guide the learning of attention based on inter-frame consistency. Instead of a global attention which aggregates information from semantically similar regions from past frames, we propose a convolutional approach

to facilitate the learning of local temporal texture, which is characteristic for reflections.

III. Temporal context for obstacle detection

Given a target frame $X \in \mathbb{R}^{3\times H\times W}$, the task of the segmentation-based obstacle detection method is to predict the segmentation mask, *i.e.* to classify each location in X as either water, sky or obstacle. We propose using the temporal context to improve the prediction accuracy. Our network (Figure 2), denoted as WaSR-T, is based on the state-of-the-art single-frame network for maritime obstacle detection WaSR [5]. We design WaSR-T to encode the discriminative temporal information about local appearance changes of a region over T preceding context frames $M \in \mathbb{R}^{T\times 3\times H\times W}$. The temporal context is added to the high-level features at the deepest level of the network as shown in Figure 2.

Following [12] and [15], the target and context frames are first individually encoded with a shared encoder network, producing per-frame feature maps $X_F \in \mathbb{R}^{N\times H\times W}$ and $M_F \in \mathbb{R}^{T\times N\times H\times W}$, where N is the number of channels. The Temporal Context Module (Section III-A) then extracts dicriminative temporal context embeddings from per-frame features. Finally, the temporal context embeddings are concatenated with target frame embeddings and fed into a decoder network. Following [5], the decoder gradually merges the information with multi-level features of the target frame (*i.e.* skip connections) and outputs the final target frame segmentation.

A. Temporal Context Module

The Temporal Context Module (TCM) extracts the temporal information from embeddings of the context and target frames and combines it with embeddings of the target frame using concatenation (Figure 2). For this reason, the number of input channels to the decoder doubles compared to the single-frame network. Thus, in order to preserve the structure and number of input channels to the decoder, TCM first reduces the dimensionality of per-frame feature maps X_F and M_F accordingly – a shared 1×1 convolutional layer is used to project the per-frame feature maps into $N/2$ dimensional per-frame embeddings X_V and M_V as shown in Figure 2.

To extract the temporal information from a sequence of frame embeddings, attention-based approaches [12], [14], [15] are often utilized, as they allow aggregation of information from semantically similar regions across multiple frames to account for movement and appearance changes of objects. Reflections, however often feature significant local texture changes as demonstrated in Figure 1. Thus, instead of globally aligning semantically similar regions using attention mechanisms, we utilize a spatio-temporal convolution to extract the local texture changes.

First we stack the context and target frame embeddings M_V and X_V into a spatio-temporal context volume $C \in \mathbb{R}^{(T+1)\times N/2\times H\times W}$. Then a 3D convolution layer is used to extract discriminative spatio-temporal features from C. To account for minor inter-frame object and camera movements, a kernel size of $(T+1)\times 3\times 3$ is used to capture temporal information in a local spatial region around locations in the context volume. We apply padding in the spatial dimensions to preserve the spatial size of the output context features $C_V \in \mathbb{R}^{N/2\times H\times W}$.

B. Efficient inference

During training, for each input image X, WaSR-T needs to extracts all per-frame context embeddings M_F in addition to target frame embeddings X_V. However, during inference the frames are passed to the network sequentially, thus recent frame embeddings can be stored in memory and feature extraction only needs to be performed on the newly observed target frame. Specifically, WaSR-T stores a buffer of T most recent frame embeddings X_V in memory and uses them as the context frame embeddings M_V in TCM. The memory buffer is initialized with T copies of the X_V embeddings of the first frame in the sequence. Using sequential inference, the efficiency of WaSR-T is not significantly impacted compared to single-frame methods, differing only due to the temporal context extraction in TCM.

IV. Experiments

A. Implementation details

WaSR-T follows the architecture of WaSR [5] and applies ResNet101 as the feature encoder. In a preliminary study we observed that in contrast to WaSR, the inertial measurements (IMU) do not bring improvements in our temporal extension. Therefore the IMU is not used in the decoder for simplicity. We apply the original WaSR training procedure, i.e., the water separation loss function, hyper-parameters, optimizers, learning rate schedule and image augmentation. We set the number of past frames in the temporal context module to $T = 5$. Because of training memory constraints, the backbone gradients are restricted to the current and previous frame. WaSR-T is trained for 100 epochs on 2×NVIDIA Titan A100S GPUs with a minibatch size of 4 per GPU.

The networks in our experiments are trained on the training set Mastr1478 (Section IV-B) and tested on the most recent maritime obstacle detection benchmark MODS [4], which contains approximately 100 annotated sequences captured under various conditions. The evaluation protocol reflects the detection performance meaningful for practical USV navigation and separately evaluates the detection of obstacle-water edge for static obstacles and the detection of dynamic obstacles. The water-edge detection robustness (μ_R) is computed from the ground truth edge, while dynamic obstacle detection is evaluated in terms of true-positive (TP), false-positive (FP) and false-negative (FN) detections, and summarized by the F1 measure, precision (Pr) and recall (Re). A dynamic obstacle counts as detected (TP) if the coverage of the segmentation inside the ground truth bounding box is sufficient, otherwise the obstacle counts as undetected (FN). Predicted segmentations outside of the ground truth bounding boxes count as false positive detections. Detection performance is reported over the entire visible navigable area and separately within a 15m *danger zone* from the USV,

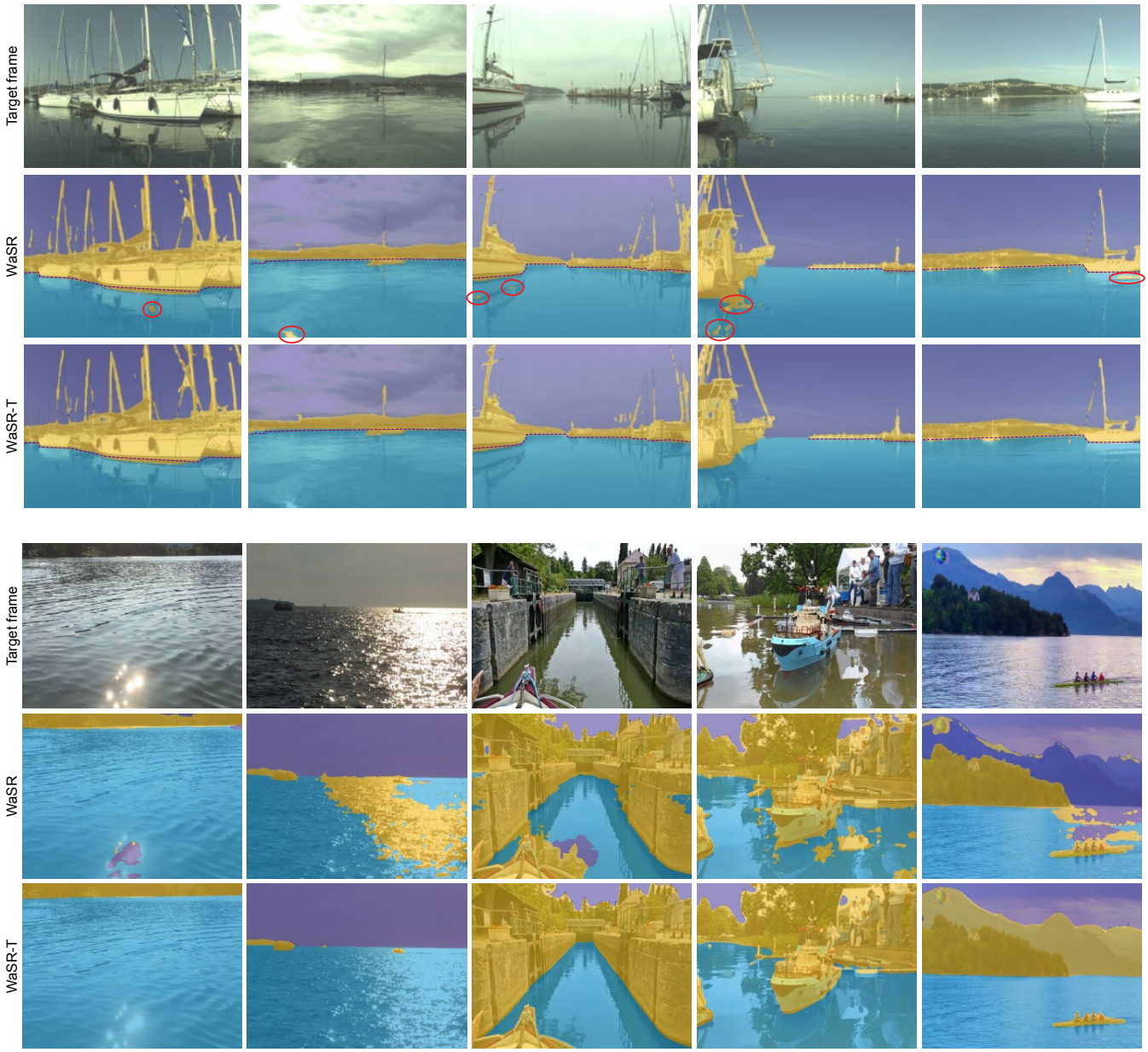

Fig. 3. Qualitative results on MODS (top) and web-sourced sequences (bottom) reveal that in WaSR-T the handling of reflections and sun glitter is significantly improved compared to WaSR, resulting in a smaller number of FP detections and increased temporal consistency.

where the detection performance is critical for immediate collision prevention.

B. Temporally extended training dataset MaSTr1478

To facilitate the training of temporal networks, we extended the recent MaSTr1325 [3] dataset, which contains 1325 fully segmented images recorded by a USV. First, the dataset was extended by adding $T = 5$ preceding frames for each annotated frame, to allow learning of the temporal context. We noticed that while MaSTr1325 is focused on the broader challenges in maritime segmentation, it contains relatively few examples of challenging reflections and glitter. We have thus extended the original dataset with additional 153 images (including their preceding frames) and use the codename *MaSTr1478* for this new dataset. The additional images were obtained from online videos or were additionally recorded by us to represent difficult scenarios for current single-frame methods, where the temporal information is important for accurate prediction, such as object mirroring, reflections and sun glitter. Examples are shown in Figure 4. The frames are labeled with per-pixel ground truth following [3]. To emphasize the challenging conditions, the training samples in the training batches are sampled from the original MaSTr1325 images and the additional images with equal probability.

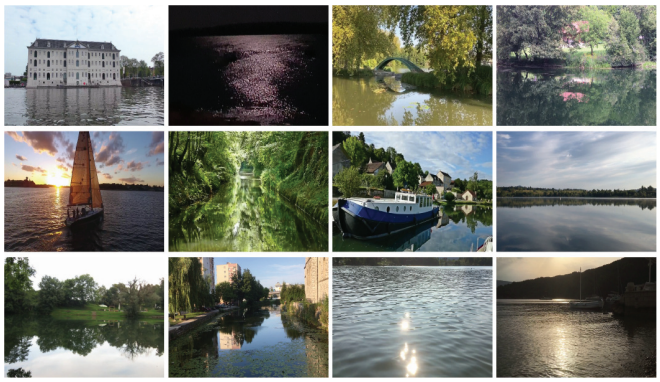

Fig. 4. Examples of the additional training sequences in the MaSTr1478 with object reflections, sun glitter and low-light conditions.

C. Comparison with state of the art

We compare WaSR-T with single-frame state-of-the-art segmentation methods (DeepLabV3+ [2], BiSeNet [17], RefineNet [18], WaSR [5]), which scored as top performers on the recent maritime obstacle detection benchmark MODS [4], as well as with state-of-the-art segmentation methods that rely on temporal information. For the latter we considered the video object segmentation method STM [12] and a recent video semantic segmentation method TMANet [15], which use memory attention to encode the temporal information from past frames. Since a relatively simple backbone is used in the original STM, we extended it to the same backbone and decoder architecture as used in WaSR [5].

Results in Table I show that multi-frame methods outperform the single-frame networks in detection precision (particularly within the danger zone), and except from TMANet, preserve a high recall. WaSR-T outperforms the original WaSR by 1.8 points in precision and 0.9 points in the overall F1, while substantially outperforming it within the danger zone resulting in a 6.0 points F1 score improvement. This is primarily due to reduction of false positives (see Figures 3 and 5), which is reflected in a 10.5 point improvement of the Pr score within the danger zone. WaSR-T also outperforms the temporal state-of-the-art networks especially inside the danger zone, resulting in approximately 2 points performance improvement of danger-zone F1 score.

In terms of speed, the new temporal module does not substantially increase the computation. The original WaSR runs at 15 FPS, while WaSR-T runs at approximately 10 FPS, which matches the sequence acquisition framerate.

Despite the large improvements in robustness to reflections, WaSR-T also shares some limitations (*e.g.* detection of thin objects) with existing methods as shown in Figure 6. For example, the temporal context is still not able to fully address reflections in rare situations where the water is completely still and the temporal texture changes cannot be observed. We aim to tackle these challenges in our future work.

TABLE I

COMPARISON OF SOTA SINGLE-FRAME AND MULTI-FRAME METHODS ON MODS IN TERMS OF WATER-EDGE DETECTION ROBUSTNESS (μ_R), PRECISION, RECALL AND F1 SCORE FOR OBSTACLE DETECTION. DANGER-ZONE PERFORMANCE IS REPORTED IN PARENTHESES.

method	μ_R	Pr	Re	F1
DeepLabV3+ [2]	96.8	80.1 (18.6)	**92.7** (**98.4**)	86.0 (31.3)
BiSeNet [17]	97.4	90.5 (53.7)	89.9 (97.0)	90.2 (69.1)
RefineNet [18]	97.3	89.0 (45.1)	93.0 (98.1)	91.0 (61.8)
WaSR [5]	97.8	95.1 (80.3)	91.9 (96.2)	93.5 (87.6)
TMANet [15]	98.3	96.4 (90.0)	85.1 (93.0)	90.4 (91.5)
STM [12]	**98.4**	96.3 (86.2)	92.5 (96.4)	**94.4** (91.0)
WaSR-T	**98.4**	**96.9** (**90.8**)	92.0 (96.5)	**94.4** (**93.6**)

D. Analysis of the alternative temporal aggregation methods

Next, we analyzed alternatives to the feature fusion in the temporal context module proposed in Section III-A: (i) pixel-wise average pooling of temporal features (window size of $T+1\times1\times1$) and (ii) local average pooling of temporal features ($T+1\times3\times3$). Table II shows that, compared to singe-frame WaSR, the simple pixel-wise temporal average pooling of context features already improves the performance over single-frame inference by 0.8 points (overall) and 1.9 points (danger zone) in F1. Increasing the pooling window size to a local window does not improve performance. In contrast, the 3D convolution approach described in Section III-A is able to learn discriminative local temporal relations and increases the F1 by an additional 0.2 points overall, and by 3.5 points inside the danger zone. The improvement is primarily on the account of substantial reduction of false positive detections.

TABLE II

WASR-T PERFORMANCE WITH DIFFERENT TEMPORAL AGGREGATION METHODS IN TERMS OF WATER-EDGE DETECTION ROBUSTNESS (μ_R), NUMBER OF FP DETECTIONS AND F1 SCORE. PERFORMANCE INSIDE THE DANGER-ZONE IS REPORTED IN PARENTHESES.

aggregation	μ_R	FP	F1
Single-frame	97.8	2492 (629)	93.5 (87.6)
Avg pool ($T+1\times1\times1$)	**98.4**	1771 (474)	94.2 (90.1)
Avg pool ($T+1\times3\times3$)	98.3	2152 (537)	93.5 (89.2)
3D Convolution	**98.4**	**1540** (**261**)	**94.4** (**93.6**)

E. Influence of the temporal and spatial context size

To gain further insights, we analyzed the influence of temporal context module parameters, i.e., the temporal context length T and spatial kernel size. Table III shows that utilizing even a single temporal context frame (i.e., $T=1$) significantly improves the performance over single-frame inference ($T=0$) by decreasing the number of false positive detections by 30% overall and 39% inside the danger zone. Increasing the temporal context length T further, brings consistent, but smaller improvements in reduction of FP detections and danger-zone F1 scores.

The spatial context size, determined by the kernel size of the 3D convolution of the temporal context module also

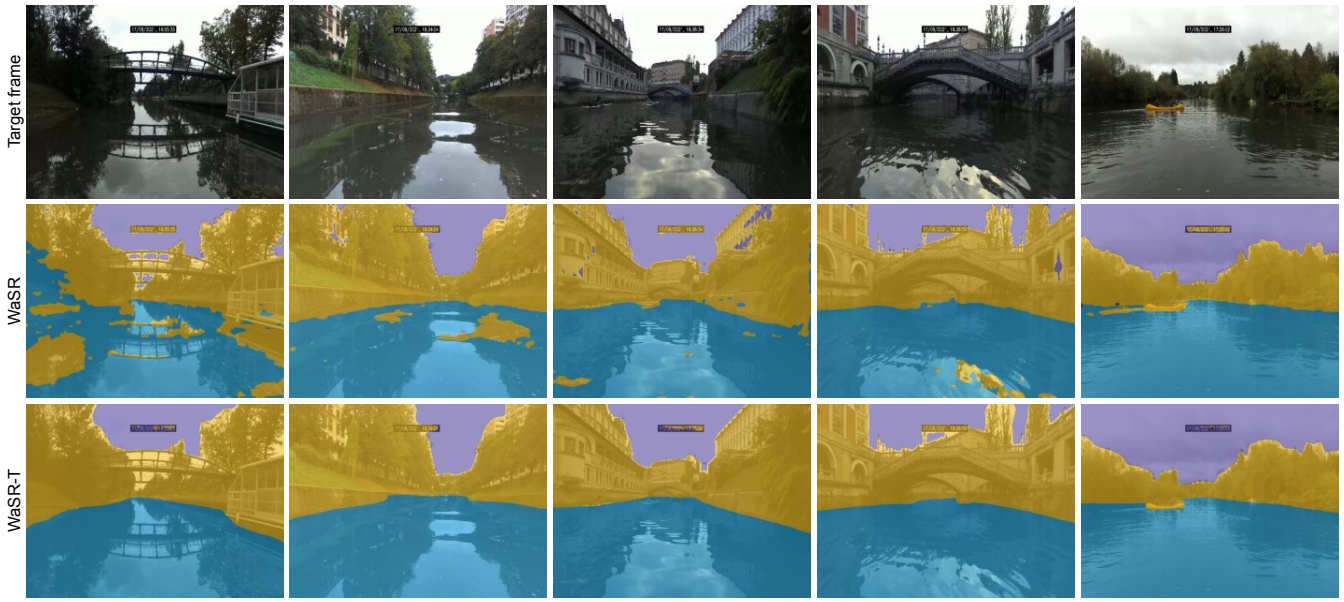

Fig. 5. Qualitative results on challenging inland water sequences demonstrates large improvements of WaSR-T in terms of practical robustness to reflections.

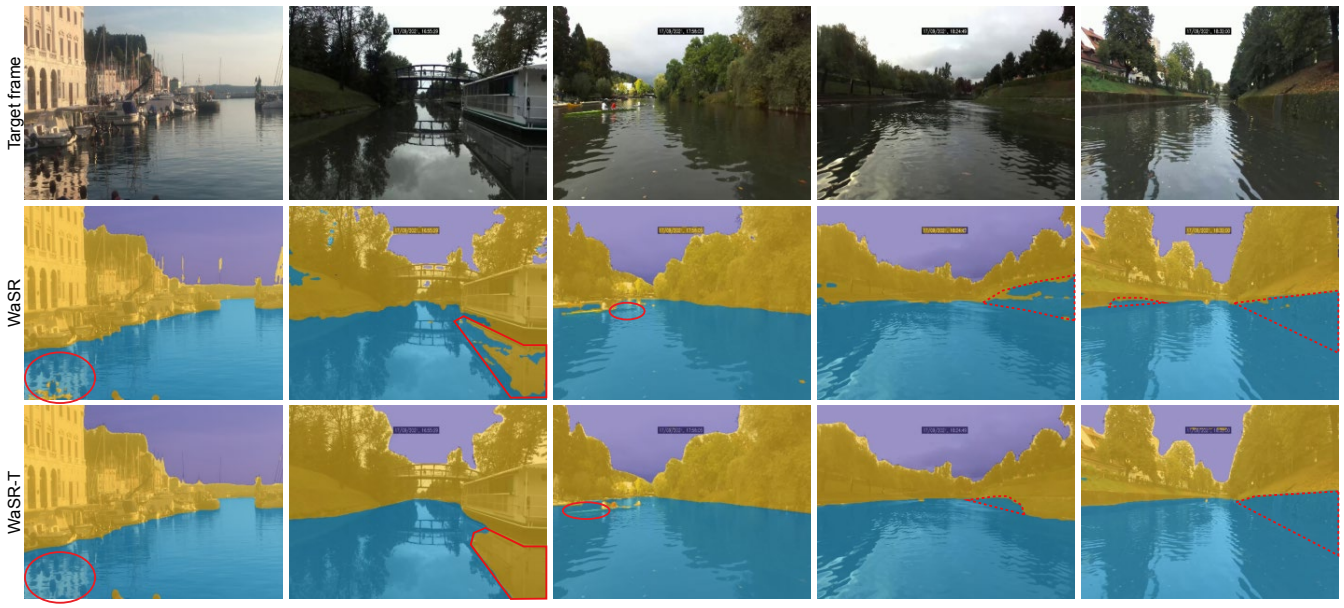

Fig. 6. Failure cases of both methods include small objects hiding in reflections (column 1), reflections on very still water (column 2), thin objects (column 3) and challenging water-land boundaries (columns 4 and 5).

importantly affects the performance. Using 1×1 spatial kernel size encodes only pixel-wise temporal relations, which negatively impacts the performance inside the danger-zone within which the objects are typically large. Increasing the kernel size to 3×3 addresses this issue, while the performance does not improve with further increasing the spatial context size.

F. Influence of the extended MaSTr1478

Finally, several experiments were performed to evaluate the contribution of the extended training dataset MaSTr1478. In particular, how much performance improvement is brought by the temporal extension and how much by the new scenes with reflections and glitter. The results in Table IV show that the single-frame WaSR does not benefit from the additional sequences in MaSTr1478. While the overall detection performance improves by 0.1 points F1, the performance decreases by 0.6 points inside the danger zone. Using only temporally extended MaSTr1325 does not improve WaSR-T performance. However, considering also the new sequences in MaSTr1478, the performance improves substantially. We observe a 41% overall reduction in the number of FP detections and a 53% reduction of FPs inside the danger zone. The overall performance is thus increased by 1.0 F1

TABLE III

INFLUENCE OF PARAMETERS IN WASR-T IN TERMS OF WATER-EDGE DETECTION ROBUSTNESS (μ_R), NUMBER OF FP DETECTIONS AND F1 SCORE. PERFORMANCE INSIDE THE DANGER-ZONE IS REPORTED IN PARENTHESES.

T	μ_R	FP	F1
0	97.8	2492 (629)	93.5 (87.6)
1	98.4	1745 (383)	94.2 (91.5)
3	**98.6**	1606 (323)	94.0 (92.6)
5	98.4	**1540** (**261**)	**94.4** (**93.6**)
kernel size			
1×1	98.1	**1456** (357)	**94.6** (92.0)
3×3	**98.4**	1540 (**261**)	94.4 (**93.6**)
5×5	98.3	1639 (318)	94.2 (92.6)

points overall and by 5.4 F1 points inside the danger zone.

Figure 3 provides qualitative results. In contrast to WaSR-T, the single-frame WaSR is unable to correctly segment regions of water containing the reflections and glitter, despite using the reflection-specific training examples of MaSTr1478. We conclude that both the new scenes and the temporal extension allow learning of the temporal appearance in WaSR-T and are responsible for improved segmentation.

TABLE IV

INFLUENCE OF TRAINING DATASET EXTENSIONS IN TERMS OF WATER-EDGE DETECTION ROBUSTNESS (μ_R), NUMBER OF FP DETECTIONS AND F1 SCORE. PERFORMANCE INSIDE THE DANGER-ZONE IS REPORTED IN PARENTHESES.

model	μ_R	FP	F1
WaSR (MaSTr1325)	97.2	2625 (561)	93.4 (88.2)
WaSR (MaSTr1478)	97.8	2492 (629)	93.5 (87.6)
WaSR-T (MaSTr1325)	97.5	2273 (655)	93.7 (87.3)
WaSR-T (MaSTr1478)	**98.4**	**1540** (**261**)	**94.4** (**93.6**)

V. CONCLUSION

We presented WaSR-T, a novel maritime obstacle detection network that harnesses the temporal context to improve obstacle detection by segmentation on water regions with ambiguous appearance. We also extended the well-known training dataset MaSTr1325 [3] by including preceding images for each training image and added new 153 training images with challenging scenes containing object mirroring and glitter – the new dataset is called MaSTr1478. Experiments show that the new images and temporal extension lead to substantial improvement on maritime obstacle detection. WaSR-T outperforms single-frame maritime obstacle detection networks as well as other networks that use temporal contexts and sets a new state-of-the-art on the maritime obstacle detection benchmark MODS [4].

REFERENCES

[1] 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 (CVPR)*, June 2016.

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Feb. 2018, pp. 801–818.

[3] B. Bovcon, J. Muhovič, J. Perš, and M. Kristan, "The MaSTr1325 dataset for training deep USV obstacle detection models," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 3431–3438.

[4] B. Bovcon, J. Muhovič, D. Vranac, D. Mozetič, J. Perš, and M. Kristan, "MODS – A USV-oriented object detection and obstacle segmentation benchmark," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–16, May 2021.

[5] B. Bovcon and M. Kristan, "WaSR–A Water Segmentation and Refinement Maritime Obstacle Detection Network," *IEEE Transactions on Cybernetics*, pp. 1–14, July 2021.

[6] T. Cane and J. Ferryman, "Evaluating deep semantic segmentation networks for object detection in maritime surveillance," in *Proceedings of AVSS 2018 - 2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2019.

[7] L. Yao, D. Kanoulas, Z. Ji, and Y. Liu, "ShorelineNet: An Efficient Deep Learning Approach for Shoreline Semantic Segmentation for Unmanned Surface Vehicles," in *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[8] L. Žust and M. Kristan, "Learning Maritime Obstacle Detection from Weak Annotations by Scaffolding," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Jan. 2022, pp. 955–964.

[9] D. Qiao, G. Liu, W. Li, T. Lyu, and J. Zhang, "Automated Full Scene Parsing for Marine ASVs Using Monocular Vision," *Journal of Intelligent & Robotic Systems*, vol. 104, no. 2, pp. 1–20, 2022.

[10] S. Varghese, S. Gujamagadi, M. Klingner, N. Kapoor, A. Bar, J. D. Schneider, K. Maag, P. Schlicht, F. Huger, and T. Fingscheidt, "An Unsupervised Temporal Consistency (TC) Loss to Improve the Performance of Semantic Segmentation Networks," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Nashville, TN, USA: IEEE, June 2021, pp. 12–20.

[11] Y. Liu, C. Shen, C. Yu, and J. Wang, "Efficient semantic video segmentation with per-frame inference," in *European Conference on Computer Vision*. Springer, 2020, pp. 352–368.

[12] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video Object Segmentation using Space-Time Memory Networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Aug. 2019, pp. 9226–9235.

[13] Y. Li, Z. Shen, and Y. Shan, "Fast Video Object Segmentation using the Global Context Module," in *European Conference on Computer Vision*. Springer, July 2020, pp. 735–750.

[14] B. Duke, A. Ahmed, C. Wolf, P. Aarabi, and G. W. Taylor, "SSTVOS: Sparse Spatiotemporal Transformers for Video Object Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5912–5921.

[15] H. Wang, W. Wang, and J. Liu, "Temporal Memory Attention for Video Semantic Segmentation," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, Sept. 2021, pp. 2254–2258.

[16] Y. Yuan, L. Wang, and Y. Wang, "CSANet for Video Semantic Segmentation With Inter-Frame Mutual Learning," *IEEE Signal Processing Letters*, vol. 28, pp. 1675–1679, 2021.

[17] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11217 LNCS, Aug. 2018, pp. 334–349.

[18] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, 2017.