НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2026 Том 26 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2026 Vol. 26 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2026-26-5-1049-1056

Synthetic dataset generation for training object detectors for underwater environment

Bassel Alshawareb[1]✉, Sergey A. Kolyubin[2]

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] balshawareb@itmo.ru✉, https://orcid.org/0009-0006-8554-0314
[2] s.kolyubin@itmo.ru, https://orcid.org/0000-0002-8057-1959

Abstract

Adapting object detection models to underwater environment presents significant challenges due to underwater light attenuation and limited data availability. This work proposes a novel approach to address the challenge of task-oriented underwater datasets scarcity by generating prompt-guided, realistic synthetic images, and subsequently fine-tuning common object detectors on this data. The proposed pipeline begins by fine-tuning a text-to-image generation model on a small dataset of real underwater images using the Low-Rank Adaptation approach. The model generates a larger set of diverse synthetic underwater images aligned with given textual prompts. To ensure that the generated images are accurate, a verification module is introduced, which filters out misaligned samples based on cosine similarity between prompt and caption embeddings of the generated image. Segmentation masks are automatically extracted from the generated images using a salient instance segmentation model. Finally, we augment real underwater images with these synthetic image-mask pairs and fine-tune a detector model on an extended dataset. Experimental results on a subset of under 800 images from the USIS10K benchmark demonstrate a 2.2 % increase in mean Average Precision (mAP) with a moderate amount of synthetic data compared to the real-only baseline. Although synthetic images are generated for only the human divers and ruins classes, detection improves also for non-augmented classes. We further experimented with a scarcer data setting by reducing the training images of the two target classes from 345 to 180. In this setting, augmenting with our synthetic images showed a larger improvement of 3.3 % in mAP, underlining the usefulness of synthetic data when real training data are limited.

Keywords

underwater object detection, Stable Diffusion XL, synthetic dataset generation, BLIP2, LoRA

For citation: Alshawareb B., Kolyubin S.A. Synthetic dataset generation for training object detectors for underwater environment. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 5, pp. 1049–1056 (in Russian). doi: 10.17586/2226-1494-2026-26-5-1049-1056

УДК 004.896

Генерация синтетических наборов данных для обучения детекторов объектов в подводной среде

Бассел Алшавареб[1]✉, Сергей Алексеевич Колюбин[2]

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] balshawareb@itmo.ru✉, https://orcid.org/0009-0006-8554-0314
[2] s.kolyubin@itmo.ru, https://orcid.org/0000-0002-8057-1959

Аннотация

Введение. Адаптация моделей обнаружения объектов к подводной среде представляет собой значительную задачу из-за ослабления света в воде и ограниченной доступности данных. Предлагается новый подход к решению проблемы нехватки целевых подводных наборов данных путем генерации управляемых текстовыми подсказками реалистичных синтетических изображений и последующей донастройки стандартных детекторов объектов на этих данных. **Метод.** Предложенный алгоритм начинается с донастройки модели генерации изображений по

© Alshawareb B., Kolyubin S.A., 2026

текстовому описанию на небольшом наборе реальных подводных изображений с использованием метода Low-Rank Adaptation. Модель генерирует большой набор разнообразных синтетических подводных изображений, соответствующих заданным текстовым подсказкам. Для обеспечения точности генерируемых изображений вводится модуль верификации, который отфильтровывает несоответствующие образцы на основе косинусного сходства между эмбеддингами подсказки и описания сгенерированного изображения. Сегментационные маски автоматически извлекаются из сгенерированных изображений с использованием модели выделения значимых объектов. В результате дополняются реальные подводные изображения синтетическими парами изображение-маска и выполняется донастройка модели детектора на расширенном наборе данных. **Основные результаты.** Экспериментальные результаты на подмножестве примерно из 800 изображений бенчмарка USIS10K показали увеличение средней точности mean Average Precision (mAP) на 2,2 % при использовании умеренного количества синтетических данных по сравнению с базовым сценарием, в котором использовались только реальные изображения. **Обсуждение.** В данном исследовании синтетические изображения сгенерированы только для классов водолазов и руин. Наблюдалось улучшение не только для исследованных классов, но и для классов, не дополненных синтетическими данными, прежде всего для рыб. Также проведены эксперименты в условиях более скудных данных, при этом сокращено число обучающих изображений двух целевых классов с 345 до 180. В этих условиях внедрение синтетических изображений дало более значительное улучшение средней точности (на 3,3 %), что подчеркивает необходимость синтетических данных при ограниченном объеме реальных обучающих данных.

Ключевые слова

обнаружение подводных объектов, Stable Diffusion XL, генерация синтетических наборов данных, BLIP2, LoRA

Ссылка для цитирования: Алшавареб Б., Колюбин С.А. Генерация синтетических наборов данных для обучения детекторов объектов в подводной среде // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 5. С. 1049–1056 (на англ. яз.). doi: 10.17586/2226-1494-2026-26-5-1049-1056

Introduction

Object detection is a critical capability for autonomous underwater robots operating in real-world environments, enabling tasks such as ecological monitoring and infrastructure inspection. Despite the advancements of object detection in terrestrial environments, known methods and models exhibit a performance drop when deployed underwater due to inherent challenges such as light scattering and absorption. This domain gap limits the reliability of Autonomous Underwater Vehicles and Remotely Operated Vehicles. Consequently, developing domain transfer approaches for models is in demand, with training and validation datasets being major enablers. Common domain adaptation methods aim to align source and target domains in the image space or feature space. Methods such as the original Domain Adaptive Faster R-CNN [1] introduce an adversarial loss to align both image-level and instance-level features between source and target domains. It integrates a domain classifier with the Faster-R-CNN framework to reduce the distribution gap. Such methods have been shown to improve mean Average Precision (mAP) on benchmarks like Foggy Cityscapes by several percentage points over source-only training [2]. Similarly, Saito et al.'s SWDA [3] adds a dual-level alignment (strong local and weak global) so that the detector learns more transferable features without changing the detector architecture.

Other methods instead directly minimize a statistical distance (e.g., maximum mean discrepancy, Jensen-Shannon divergence, or CORrelation ALignment [4]) between source and target feature distributions. By introducing extra loss terms (and sometimes additional modules) into a standard detector like Faster-R-CNN or SSD, these approaches guide the network to learn representations that are less sensitive to domain shifts.

Domain Adaptive DEtection TRansformer (DA-DETR) [5] has extended the DETR framework by integrating an adversarial loss and a feature fusion module into the DETR architecture.

Instead of — or in addition to — adapting features with unlabeled target images, researchers have explored the creation of synthetic datasets that closely mimic the target domain. These synthetic datasets serve as the downstream training dataset when an annotated target dataset is not available, and a supplementary dataset when it is available. Image-to-image translation methods are used to generate training data that cover a wide range of target conditions. An example is the work by Borrego et al. [6], where even non-photorealistic synthetic images are randomized (e.g., in texture, lighting, and color). Other works allow for training on fully synthetic data. For instance, [7] uses CycleGAN to generate realistic images from computer-aided design 3D models, while [8] adapts models that are trained on synthetic data to real domain by generating another synthetic more realistic data and fine-tuning on it. Similarly, [9] generates an intermediate synthetic dataset and subsequently fine-tunes an object detector on it along with the target dataset. More recent work on Unsupervised Domain Adaptation (UDA) for anchorless detectors has shown that UDA methods can be incorporated to adapt a detector trained on synthetic images to real images [10].

Recently, generative models like Stable Diffusion XL [11] and Flux.1[1] demonstrated a remarkable capability for generating realistic synthetic data. By utilizing a relatively small amount of data, it is possible to fine-tune the model to generate images from the same distribution adapted for the domain of interest, which is underwater in our case. A recent diffusion-based approach generates synthetic adverse weather images (fog, rain, and snow) using text-to-image models like Midjourney [12]. Their synthetic data are designed to match the camera viewpoints, scene composition, and weather conditions of the target

[1] Black Forest Labs. Flux Available at: https://github.com/black-forest-labs/flux (accessed: 01.09.2026).

domain (e.g., Foggy Cityscapes). Other works apply image-to-image translation using instruct pix2pix to generate data that simulates the target domain, while preserving the location information of the objects [13]. However, generative models that are conditioned on images require access to the source images, which might be unavailable, or available in limited amounts. On the other hand, text-to-image approaches do not require access to source images, but they need to be manually labeled with bounding boxes for object detection, which can be inefficient with a large amount of generated data.

This study proposes a novel pipeline utilizing synthetic data generated with Stable Diffusion XL for adapting detectors to underwater environments. Unlike related prior methods, our framework does not require access to source domain data or manual bounding box annotations of synthetic images. Our approach treats underwater image generation as full-scene text-to-image synthesis, training the generator on entire underwater images.

Related Work

Researchers have adopted different approaches to improve underwater object detection by addressing the domain shift, which stems primarily from light scattering and absorption, and the nature of underwater objects.

The main line of work on underwater object detection is centred around applying image enhancement methods as a preprocessing step prior to detection. In contrast to domain adaptation methods, image enhancement methods aim to convert target images to fit the distribution of the source domain by learning the challenging underwater lighting conditions and removing them from the image. Whether using a dedicated image enhancement model to learn underwater conditions [14] or through joint enhancer-detector training [15], [16], these methods often introduce artifacts, potentially causing loss of important details of objects, which harms object detection performance. Therefore, other works train detectors directly on underwater images without the need for image enhancement. Such approaches often introduce specialized modules to deal with the specific nature of underwater images, like degraded quality [16], small and occluded objects [17], or color distortion [18]. Despite their effectiveness, these solutions are often designed for a specific model architecture, limiting their application in different tasks.

The mentioned methods attempt to fix the input or the model, but do not address the root issue: limited training data that captures underwater conditions. Motivated by that, other methods focus on augmenting training data with synthetic underwater data that simulate real underwater conditions. Different methods were proposed to obtain synthetic data such as using domain-specific simulators [19], web scraping [20], or image generation models [21]. Among these methods, generative models revealed outstanding capabilities at generating photorealistic images, making them a promising candidate towards synthetic underwater images with realistic lighting conditions.

Patricio et al. [21] fine-tune Flux.1 using Low-Rank Adaptation (LoRA) [22] on foreground objects and generate a synthetic training dataset for object detection. However, in their approach the generator is trained only on cropped objects from the underwater image, with the generated objects appearing in fixed spatial locations after image composition. This scheme deprives the generator of background context, which can contain important information needed for generating physically consistent images. Furthermore, composing generated objects onto fixed backgrounds limits the diversity of object placements, sizes, and poses.

To address these limitations, our work instead exposes the generator to the entire image under text-to-image generation setting, while obtaining ground truth masks automatically using a zero-shot salient instance segmentation model. We hypothesize that full-scene generation better captures environmental effects, such as underwater light attenuation, leading to synthetic data that more faithfully simulates real underwater conditions and can therefore benefit the detection of underwater objects beyond the specific classes that are augmented.

Method

This work proposes a three-stage pipeline for generating a synthetic underwater dataset for salient instance segmentation, illustrated in Fig. 1. Stage 1 fine-tunes a text-to-image generator on real underwater data. Stage 2 generates images with automatic prompt-alignment filtering followed by manual visual curation. Stage 3 applies automatic salient instance segmentation to the curated images.

In the first stage, we fine-tune the generator using LoRA on underwater images of the target dataset with their captions. For each selected attention layer l, LoRA freezes the pre-trained weights W_l and introduces a trainable low-rank update ΔW_l scaled by a hyperparameter α_l. The updated weights are computed as:

$$W_l' = W_l + \alpha_l \Delta W_l,$$

where $\Delta W_l = B_l A_l$, with $B_l \in \mathbb{R}^{d \times r}$ and $A_l \in \mathbb{R}^{r \times k}$ being the low-rank matrices, and $r << \min(d, k)$. This formulation enables efficient adaptation to the underwater domain with substantially fewer trainable parameters compared to full fine-tuning. Similar to [23], we generate image captions using BLIP2 [24] to obtain text-image pairs for fine-tuning the generator. In the second stage, we use the fine-tuned model to generate images of a target class. In this work, we generate images of ruins and human divers classes only. We include in the prompt words like “low visibility”, or “poor quality” to generate images with hard conditions. The generated images do not always depict the object of interest and may contain visual artifacts. Therefore, we introduce a verification block that ensures alignment between the generated image and the input prompt. The verification block computes the cosine similarity between sentence embeddings of the input prompt and the embeddings of the caption of the generated image, and rejects images whose similarity falls below a threshold tuned manually. In this case, the generation is repeated with a different random seed. The successfully generated images proceed to stage 3,

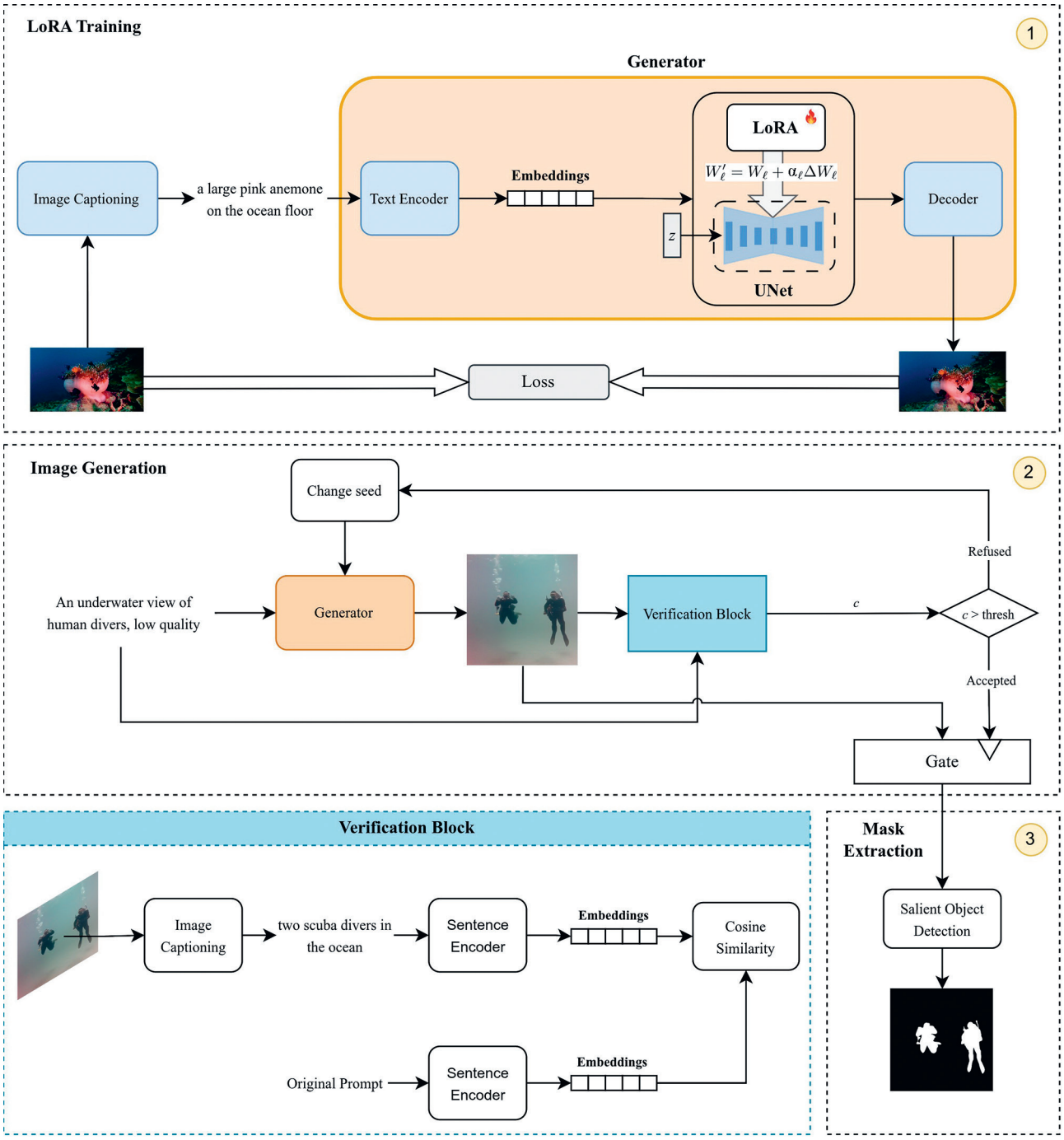


Fig. 1. The proposed three-stage pipeline for synthetic dataset generation. z indicates the random seed used for generation and c indicates the cosine similarity score

where InSPyReNet [25], a salient instance segmentation model, is used to extract ground truth masks. Finally, the generated images with their masks are added to the real training set to fine-tune object detection models.

Experiments

We conduct all experiments on the underwater USIS10K dataset [18]. The original dataset contains approximately 7500 training images and 1500 validation/test images across seven categories: ruins, fish, reefs, aquatic plants, human divers, robots, and sea-floor. In the experiments we first selected a subset of 1000 training images from the USIS10K dataset, with ruins and human divers classes having the least number of objects, reflecting the inherent data imbalance in the USIS10K dataset. During inspection of the USIS10K images, we identified multiple issues with the original data labeling (specifically human divers class): misclassified annotations, imprecise bounding boxes with at least twice the size of the object, and objects of incorrect saliency labeling (salient objects omitted or non-salient objects annotated). We corrected annotations of human divers class in the test split to ensure reliable evaluation. For aquatic plants and sea-floor, accurate

annotation requires underwater domain expertise that was not consistently applied; we therefore removed these two classes entirely from our study. In addition, we found that the annotations of the robots class contain instruments and other objects beside robots, so this class was also removed. The obtained train split contains 130 ruins, 222 fish, 220 reefs, and 215 human divers real images, with fish and reefs having a substantially larger number of objects compared to ruins and human divers classes.

In the conducted experiments, we fine-tune Stable Diffusion XL to generate 1000 prompt-aligned images for each of human divers and ruins classes. Then we manually inspect and remove generated images that show redundant contents. We keep 350 images for each class for our experiments. We ensure the selected images are free of noticeable artifacts, and that they show the target class as the only salient object in the image. For the ruins class, we include specific subclasses in the prompts, such as wrecks and statues. We create three setups varying the number of synthetic images used: no synthetic images (baseline), synthetic images added until object count is balanced among the classes (moderate), all synthetic images. We then train the Mask R-CNN detector [26] on the augmented data for 50 epochs using stochastic gradient descent with a learning rate of 0.008, momentum of 0.9, and weight decay of 0.0001. We repeat every training experiment of Mask R-CNN four times with different random seeds and record the mean and standard deviation of the mAP@[.50:.95] on our reannotated test set. All object detection experiments were conducted on a single NVIDIA RTX 5090 GPU, and the generation experiments on a single NVIDIA RTX 4090 GPU. Using a moderate amount of our synthetic data, we obtained a mean improvement of 2.2 % in mAP over the real-only baseline (Table 1).

To understand the contribution of synthetic images in data-scarce scenarios, we select a smaller subset containing 100 human divers and 80 ruins real images. The removed images also contained 38 reef and 26 fish object instances, so this reduction additionally lowers the amount of real data available for the non-augmented classes. We repeat the same procedure used earlier to create three training setups (Table 2). We obtain a more pronounced improvement of 3.3 % mAP in the moderate setup. More synthetic images further increase AP for the human divers class but decrease

Table 1. Mean Average Precision (mAP@[.50:.95]) of models trained on the selected subset of the USIS10K dataset with varying number of added synthetic images

Setup	Num Synth		AP				
	Ruins	Divers	Ruins	Fish	Reefs	Divers	Avg
USIS-SAM	0	0	24.6	44.3	17.7	52.3	34.7
Subset	0	0	17.7±1.0	29.9 ± 1.3	9.1 ± 0.4	46.3 ± 1.1	25.7 ± 0.8
Subset balanced	100	250	**21.1 ± 1.3**	31.4 ± 0.9	**9.4 ± 0.6**	49.7 ± 1.4	27.9 ± 0.3
Subset + all synth	350	350	20.8 ± 1.1	**31.5 ± 1.3**	8.4 ± 0.5	**52.0 ± 0.6**	**28.2 ± 0.3**

Note. The first row shows USIS-SAM, the state-of-the-art model, trained on the real-only subset. The rest of the rows show experiments using Mask R-CNN. The table displays per-class Average Precision (AP) as well as the overall mean average precision (Avg). Highest mean values of Mask R-CNN experiments are marked in bold. The column Num Synth indicates the number of synthetic images added to the training data.

Table 2. Mean average precision (mAP@[.50:.95]) of Mask R-CNN on the scarce-data subset of the USIS10K dataset with different numbers of added synthetic images

Setup	Num Synth		AP				
	Ruins	Divers	Ruins	Fish	Reefs	Divers	Avg
small	0	0	12.5 ± 1.0	29.3 ± 0.4	7.9 ± 0.1	38.9 ± 1.7	22.1 ± 0.6
small balanced	150	250	**15.6 ± 1.1**	31.3 ± 0.4	**8.5 ± 0.1**	46.2 ± 0.8	25.4 ± 0.4
small + all synth	350	350	14.4 ± 0.6	**31.5 ± 0.6**	**8.5 ± 0.0**	**48.2 ± 0.5**	**25.6 ± 0.3**

Table 3. Per-class average recall (AR@[.50:.95]) of Mask R-CNN on the selected USIS10K subset and the scarce-data subset, with varying numbers of added synthetic images

Setup	Ruins	Fish	Reefs	Divers	AR
Subset	30.1 ± 3.5	36.2 ± 1.8	21.9 ± 2.5	55.8 ± 1.3	44.2 ± 1.9
Subset balanced	**34.3 ± 1.8**	**38.4 ± 2.8**	**23.2 ± 3.1**	60.3 ± 1.1	**46.4 ± 3.3**
Subset + all synth	32.5 ± 1.3	37.0 ± 1.2	20.9 ± 1.4	**62.0 ± 0.3**	44.0 ± 0.9
small	20.6 ± 1.5	34.4 ± 0.6	19.2 ± 0.9	47.4 ± 1.0	40.8 ± 0.4
small balanced	**24.8 ± 1.2**	36.5 ± 0.4	19.9 ± 0.7	55.5 ± 1.1	42.3 ± 0.2
small + all synth	24.5 ± 1.4	**37.4 ± 1.0**	**20.8 ± 1.0**	**57.7 ± 0.7**	**42.8 ± 0.7**

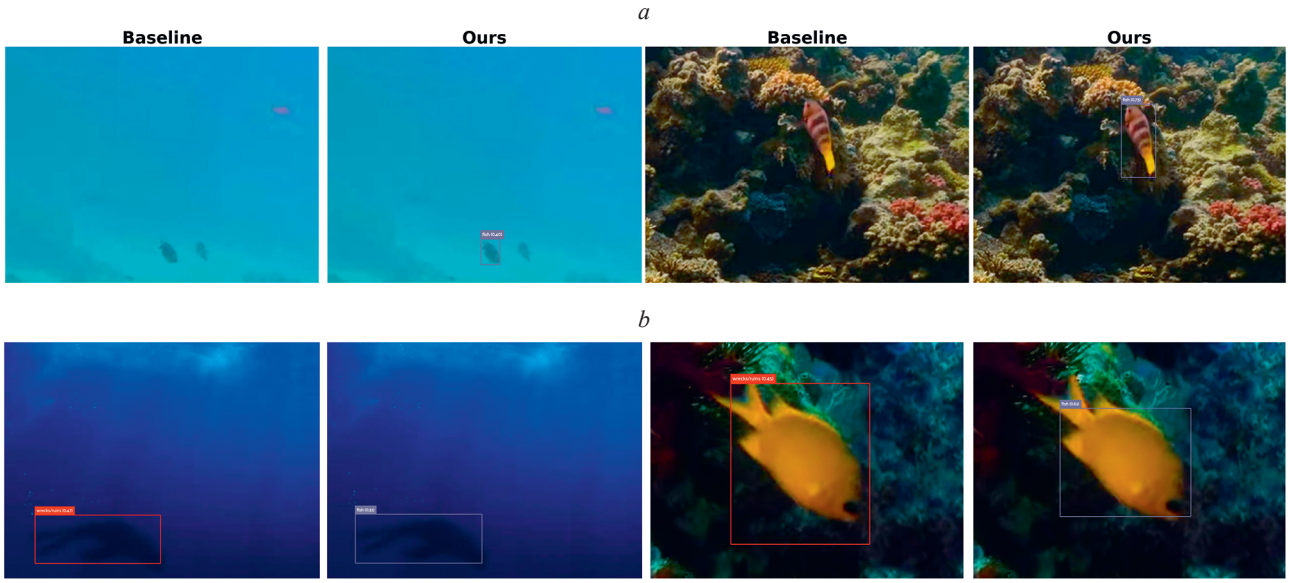


Fig. 2. Qualitative fish-class detection results for the baseline (real images only) and the model trained with synthetic divers and ruins images: baseline false negatives (*a*); baseline misclassifications (*b*)

it for ruins. We attribute this to limited object diversity in the synthetic ruins images, to which the detector may overfit.

Notably, the improvement is not confined to the augmented classes: fish detection also improves by 1.5–2.2 AP in different setups, whereas reefs improve marginally. Because synthetic images are added only for human divers and ruins, these cross-class gains may indicate that full-scene synthetic data also convey general underwater appearance cues, such as light attenuation, in addition to class-specific labels.

The cross-class effect is also evident in recall (Table 3), indicating that the augmented detector finds more objects in real underwater images. Qualitative fish-class examples in Fig. 2 show fewer missed detections and misclassifications compared with the baseline.

Conclusion

This work introduces a pipeline for creating synthetic instance segmentation datasets tailored to the underwater domain using image generation, image captioning, and salient instance segmentation models. It ensures full exposure of the generator to the entire training image based on text-to-image training. Our main finding is that generating full-scene synthetic images for only two classes (human divers and ruins) improves detection on those classes and also yields measurable gains for non-augmented classes, most consistently for fish. This suggests that our synthetic images capture domain-specific visual characteristics that help the detector adapt to real underwater conditions. Our experiments show a mean Average Precision gain of 2.2 % on a subset of USIS10K with moderate synthetic data and a 3.3 % gain in a scarce-data setting. Although our pipeline automatically produces segmentation masks for generated images, manual curation is still required to remove artifacts and repetitive patterns. Future work includes developing quality metrics for systematic sample selection and extending the synthetic dataset to additional classes.

References

1. Chen Y., Li W., Sakaridis C., Dai D., Van Gool L. Domain adaptive Faster R-CNN for object detection in the wild. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3339–3348. doi: 10.1109/CVPR.2018.00352
2. Sakaridis C., Dai D., Van Gool L. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 2018, vol. 126, no. 9, pp. 973–992. doi: 10.1007/s11263-018-1072-8
3. Saito K., Ushiku Y., Harada T., Saenko K. Strong-weak distribution alignment for adaptive object detection. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6949–6958. doi: 10.1109/CVPR.2019.00712
4. Sun B., Saenko K. Deep CORAL: Correlation alignment for deep domain adaptation. *Lecture Notes in Computer Science*, 2016, vol. 9915, pp. 443–450. doi: 10.1007/978-3-319-49409-8_35
5. Zhang J., Huang J., Luo Z., Zhang G., Zhang X., Lu S. DA-DETR: Domain Adaptive Detection Transformer with Information Fusion.

Литература

1. Chen Y., Li W., Sakaridis C., Dai D., Van Gool L. Domain adaptive Faster R-CNN for object detection in the wild // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 3339–3348. doi: 10.1109/CVPR.2018.00352
2. Sakaridis C., Dai D., Van Gool L. Semantic foggy scene understanding with synthetic data // International Journal of Computer Vision. 2018. V. 126. N 9. P. 973–992. doi: 10.1007/s11263-018-1072-8
3. Saito K., Ushiku Y., Harada T., Saenko K. Strong-weak distribution alignment for adaptive object detection // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. P. 6949–6958. doi: 10.1109/CVPR.2019.00712
4. Sun B., Saenko K. Deep CORAL: Correlation alignment for deep domain adaptation // Lecture Notes in Computer Science. 2016. V. 9915. P. 443–450. doi: 10.1007/978-3-319-49409-8_35
5. Zhang J., Huang J., Luo Z., Zhang G., Zhang X., Lu S. DA-DETR: Domain Adaptive Detection Transformer with Information Fusion //

Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23787–23798. doi: 10.1109/CVPR52729.2023.02278
6. Borrego J., Dehban A., Figueiredo R., Moreno P., Bernardino A., Santos-Victor J. Applying domain randomization to synthetic data for object category detection. *arXiv*, 2018. arXiv:1807.09834. doi: 10.48550/arXiv.1807.09834
7. Hu J., Xiao F., Jin Q., Zhao G., Lou P. Synthetic data generation based on RDB-CycleGAN for industrial object detection. *Mathematics*, 2023, vol. 11, no. 22, pp. 4588. doi: 10.3390/math11224588
8. Brehm S., Scherer S., Lienhart R. Semantically consistent image-to-image translation for unsupervised domain adaptation. *Proc. of the 14th International Conference on Agents and Artificial Intelligence*, 2022, vol. 2, pp. 131–141. doi: 10.5220/0010786000003116
9. Hsu H.-K., Yao C.-H., Tsai Y.-H., Hung W.-C., Tseng H.-Y., Singh M., et al. Progressive domain adaptation for object detection. *arXiv*, 2019. arXiv:1910.11319. doi: 10.48550/arXiv.1910.11319
10. Scheck T., Grassi A.P., Hirtz G. Unsupervised domain adaptation from synthetic to real images for anchorless object detection. *Proc. of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2021, vol. 5, pp. 319–327. doi: 10.5220/0010202503190327
11. Podell D., English Z., Lacey K., Blattmann A., Dockhorn T., Müller J., et al. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv*, 2023. arXiv:2307.01952. doi: 10.48550/arXiv.2307.01952
12. Rothmeier T., Huber W., Knoll A.C. Time to shine: fine-tuning object detection models with synthetic adverse weather images. *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 4435–4444. doi: 10.1109/wacv57701.2024.00439
13. Gurbindo U., Brando A., Abella J., König C. Object detection in adverse weather conditions for autonomous vehicles using Instruct Pix2Pix. *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8. doi: 10.1109/ijcnn64981.2025.11228623
14. Liu R., Jiang Z., Yang S., Fan X. Twin adversarial contrastive learning for underwater image enhancement and beyond. *IEEE Transactions on Image Processing*, 2022, vol. 31, pp. 4922–4936. doi: 10.1109/TIP.2022.3190209
15. Chang L., Wang Y., Deng L., Du B., Xu C. WaterDiffusion: Learning a prior-involved unrolling diffusion for joint underwater saliency detection and visual restoration. *Proc. of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, no. 2, pp. 1998–2006. doi: 10.1609/aaai.v39i2.32196
16. Lian S., Li H., Cong R., Li S., Zhang W., Kwong S. WaterMask: Instance segmentation for underwater imagery. *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1305–1315. doi: 10.1109/ICCV51070.2023.00126
17. Li Z., Zheng B., Chao D., Zhu W., Li H., Duan J., et al. Underwater-Yolo: Underwater object detection network with dilated deformable convolutions and dual-branch occlusion attention mechanism. *Journal of Marine Science and Engineering*, 2024, vol. 12, no. 12, pp. 2291. doi: 10.3390/jmse12122291
18. Lian S., Zhang Z., Li H., Li W., Yang L.T., Kwong S., et al. Diving into underwater: segment anything model guided underwater salient instance segmentation and a large-scale dataset. *Proc. of the 41st International Conference on Machine Learning*, 2024, pp. 29545–29559
19. Xia Y., Deng S., Li Y., Zuo W., Xin J., Zheng N. SUNOD: Synthetic underwater non-natural object detection dataset. *Proc. of the China Automation Congress (CAC)*, 2023, pp. 5577–5582. doi: 10.1109/CAC59555.2023.10450530
20. Nandyala N.R., Sanodiya R.K. Underwater object detection using synthetic data. *Proc. of the 11th International Symposium on Electronic Systems Devices and Computing (ESDC)*, 2023, pp. 1–6. doi: 10.1109/esdc56251.2023.10149870
21. Patrício A., Dehban A., Ventura R. FLORA: Efficient synthetic data generation for object detection in low-data regimes via finetuning Flux LoRA. *arXiv*, 2025. arXiv:2508.21712v1. doi: 10.48550/arXiv.2508.21712
22. Hu E.J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., et al. LoRA: Low-rank adaptation of large language models. *arXiv*, 2021. arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685
23. Zhang F., You S., Li Y., Fu Y. Atlantis: enabling underwater depth estimation with stable diffusion. *Proc. of the IEEE/CVF Conference*

Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. P. 23787–23798. doi: 10.1109/CVPR52729.2023.02278
6. Borrego J., Dehban A., Figueiredo R., Moreno P., Bernardino A., Santos-Victor J. Applying domain randomization to synthetic data for object category detection // arXiv. 2018. arXiv:1807.09834. doi: 10.48550/arXiv.1807.09834
7. Hu J., Xiao F., Jin Q., Zhao G., Lou P. Synthetic data generation based on RDB-CycleGAN for industrial object detection // Mathematics. 2023. V. 11. N 22. P. 4588. doi: 10.3390/math11224588
8. Brehm S., Scherer S., Lienhart R. Semantically consistent image-to-image translation for unsupervised domain adaptation // Proc. of the 14th International Conference on Agents and Artificial Intelligence. 2022. V. 2. P. 131–141. doi: 10.5220/0010786000003116
9. Hsu H.-K., Yao C.-H., Tsai Y.-H., Hung W.-C., Tseng H.-Y., Singh M., et al. Progressive domain adaptation for object detection // arXiv. 2019. arXiv:1910.11319. doi: 10.48550/arXiv.1910.11319
10. Scheck T., Grassi A.P., Hirtz G. Unsupervised domain adaptation from synthetic to real images for anchorless object detection // Proc. of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. 2021. V. 5. P. 319–327. doi: 10.5220/0010202503190327
11. Podell D., English Z., Lacey K., Blattmann A., Dockhorn T., Müller J., et al. SDXL: Improving latent diffusion models for high-resolution image synthesis // arXiv. 2023. arXiv:2307.01952. doi: 10.48550/arXiv.2307.01952
12. Rothmeier T., Huber W., Knoll A.C. Time to shine: fine-tuning object detection models with synthetic adverse weather images // Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2024. P. 4435–4444. doi: 10.1109/wacv57701.2024.00439
13. Gurbindo U., Brando A., Abella J., König C. Object detection in adverse weather conditions for autonomous vehicles using Instruct Pix2Pix // Proc. of the International Joint Conference on Neural Networks (IJCNN). 2025. P. 1–8. doi: 10.1109/ijcnn64981.2025.11228623
14. Liu R., Jiang Z., Yang S., Fan X. Twin adversarial contrastive learning for underwater image enhancement and beyond // IEEE Transactions on Image Processing. 2022. V. 31. P. 4922–4936. doi: 10.1109/TIP.2022.3190209
15. Chang L., Wang Y., Deng L., Du B., Xu C. WaterDiffusion: Learning a prior-involved unrolling diffusion for joint underwater saliency detection and visual restoration // Proc. of the AAAI Conference on Artificial Intelligence. 2025. V. 39. N 2. P. 1998–2006. doi: 10.1609/aaai.v39i2.32196
16. Lian S., Li H., Cong R., Li S., Zhang W., Kwong S. WaterMask: Instance segmentation for underwater imagery // Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023. P. 1305–1315. doi: 10.1109/ICCV51070.2023.00126
17. Li Z., Zheng B., Chao D., Zhu W., Li H., Duan J., et al. Underwater-Yolo: Underwater object detection network with dilated deformable convolutions and dual-branch occlusion attention mechanism // Journal of Marine Science and Engineering. 2024. V. 12. N 12. P. 2291. doi: 10.3390/jmse12122291
18. Lian S., Zhang Z., Li H., Li W., Yang L.T., Kwong S., et al. Diving into underwater: segment anything model guided underwater salient instance segmentation and a large-scale dataset // Proc. of the 41st International Conference on Machine Learning. 2024. P. 29545–29559
19. Xia Y., Deng S., Li Y., Zuo W., Xin J., Zheng N. SUNOD: Synthetic underwater non-natural object detection dataset // Proc. of the China Automation Congress (CAC). 2023. P. 5577–5582. doi: 10.1109/CAC59555.2023.10450530
20. Nandyala N.R., Sanodiya R.K. Underwater object detection using synthetic data // Proc. of the 11th International Symposium on Electronic Systems Devices and Computing (ESDC). 2023. P. 1–6. doi: 10.1109/esdc56251.2023.10149870
21. Patrício A., Dehban A., Ventura R. FLORA: Efficient synthetic data generation for object detection in low-data regimes via finetuning Flux LoRA // arXiv. 2025. arXiv:2508.21712v1. doi: 10.48550/arXiv.2508.21712
22. Hu E.J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., et al. LoRA: Low-rank adaptation of large language models // arXiv. 2021. arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685
23. Zhang F., You S., Li Y., Fu Y. Atlantis: enabling underwater depth estimation with stable diffusion // Proc. of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 11852–11861. doi: 10.1109/cvpr52733.2024.01126
24. Li J., Li D., Savarese S., Hoi S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *Proc. of the 40th International Conference on Machine Learning*, 2023, pp. 19730–19742.
25. Kim T., Kim K., Lee J., Cha D., Lee J., Kim D. Revisiting image pyramid structure for high resolution salient object detection. *Lecture Notes in Computer Science*, 2023, vol. 13847, pp. 257–273. doi: 10.1007/978-3-031-26293-7_16
26. He K., Gkioxari G., Dollár P., Girshick R. Mask R-CNN. Proc. of the *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988. doi: 10.1109/ICCV.2017.322

on Computer Vision and Pattern Recognition (CVPR). 2024. P. 11852–11861. doi: 10.1109/cvpr52733.2024.01126
24. Li J., Li D., Savarese S., Hoi S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models // Proc. of the 40th International Conference on Machine Learning. 2023. P. 19730–19742.
25. Kim T., Kim K., Lee J., Cha D., Lee J., Kim D. Revisiting image pyramid structure for high resolution salient object detection // Lecture Notes in Computer Science. 2023. V. 13847. P. 257–273. doi: 10.1007/978-3-031-26293-7_16
26. He K., Gkioxari G., Dollár P., Girshick R. Mask R-CNN // Proc. of the IEEE International Conference on Computer Vision (ICCV). 2017. P. 2980–2988. doi: 10.1109/ICCV.2017.322

Authors

Bassel Alshawareb — Engineer, PhD student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0006-8554-0314, balshawareb@itmo.ru
Sergey A. Kolyubin — D.Sc., Associate Professor, Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 35303066700, https://orcid.org/0000-0002-8057-1959, s.kolyubin@itmo.ru

Авторы

Алшавареб Бассел — инженер, аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0006-8554-0314, balshawareb@itmo.ru
Колюбин Сергей Алексеевич — доктор технических наук, доцент, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 35303066700, https://orcid.org/0000-0002-8057-1959, s.kolyubin@itmo.ru

Received 08.05.2026
Approved after reviewing 25.07.2026
Accepted 19.09.2026

Статья поступила в редакцию 08.05.2026
Одобрена после рецензирования 25.07.2026
Принята к печати 19.09.2026

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»