조건부 Diffusion Model 기반 전처리 및 증강을 통한 수중 객체 탐지 향상 연구

순동현[1*], 이태복[2*]
대구경북과학기술원[1], 국민대학교[2]
dhsoon@dgist.ac.kr[1], plkj3078@kookmin.ac.kr[2]

Enhancing Underwater Object Detection via Conditional Diffusion Model-based Preprocessing and Augmentation

Donghyeon Soon[1*], Taebok Lee[2*]
School of Undergraduate of Studies, DGIST[1], School of Electrical Engineering, Kookmin University[2]

요 약

수중 객체 탐지는 낮은 가시성, 색 왜곡, 환경적 노이즈뿐 아니라 고품질 영상 확보의 어려움과 잦은 품질 저하(흐림, 저해상도, 색 편향)로 인해, 딥러닝 기반 모델이 데이터 부족과 일반화 성능 한계에 직면하는 까다로운 과제이다. 본 연구는 이를 해결하기 위해 조건부 확산 기반 전처리와 자동 증강 파이프라인을 제안한다. 전처리 단계에서는 이미지 기반 조건을 활용해 색 왜곡과 노이즈를 보정하고, 증강 단계에서는 텍스트 조건과 IP-Adapter 를 이용해 객체는 보존한 채 다양한 배경을 합성함으로써 추가적인 라벨링 없이 대규모 학습 데이터 확보를 가능하게 한다. 또한 포토메트릭 증강을 병행하여 수중 영상 특유의 색 왜곡을 완화하였다. 실험 결과, 제안 기법은 Baseline 대비 mAP@50 약 6%p, mAP@[0.50:0.95] 약 8%p 향상을 보였다. 각 기법은 개별적으로도 유의미한 향상을 보였으며, 결합 시 상호 보완적 효과를 통해 추가적인 성능 향상을 달성하였다. 본 연구는 데이터 확보가 어려운 다양한 비전 분야에도 적용 가능한 효율적 증강 전략을 제시한다.

Ⅰ. 서 론

수중 객체 탐지는 낮은 가시성, 색 왜곡, 환경적 노이즈 등 복합적인 요인으로 인해 육상 환경보다 훨씬 다루기 어려운 과제로 인식되고 있다. 특히, 고품질의 수중 영상을 대규모로 확보하기 어렵고, 확보된 데이터조차 흐림·저해상도·색 편향과 같은 품질 저하 문제가 빈번히 발생한다. 이러한 제약으로 인해 딥러닝 기반 탐지 모델은 일반화 성능이 크게 제한되며, 결국 수중 객체 탐지 분야는 데이터 부족과 품질 저하라는 이중적인 한계에 직면해 있다. 이러한 문제를 동시에 해결하기 위해, 본 연구에서는 조건부 확산 기반 전처리 및 자동 증강 파이프라인을 제안한다. 제안된 방법은 원본 데이터의 품질을 개선하는 전처리 단계와, 객체 주석을 유지한 채 다양한 수중 환경을 합성하는 증강 단계로 구성된다. 이를 통해 데이터 품질 저하와 부족 문제를 동시에 완화하고, 탐지 모델의 학습 효율성과 일반화 성능을 향상시키는 것을 목표로 한다. 본 논문은 제안된 파이프라인을 YOLOv8[1] 모델에 적용하여 성능 향상을 입증하고, 구성 요소별 검증 실험을 통해 각 기법의 효과를 체계적으로 분석한다.

Ⅱ. 본론

2.1 배경지식

확산 모델[2](Diffusion Model)은 데이터 x_0에 점진적으로 잡음을 더하는 정방향 과정 $q(x_t \mid x_{t-1})$과, 잡음이 섞인 x^t 로부터 원신호를 복원하는 역과정 $p_\theta(x_{t-1} \mid x_t, c)$ 을 학습하여 이미지를 생성한다. 여기서 c는 텍스트, 깊이맵, 참조 이미지 등 다양한 조건(condition)이 될 수 있으며, 조건부확산(Conditional Diffusion)은 이러한 조건을 활용해 원하는 방향으로 합성을 제어한다.
Stable Diffusion[3]은 대표적인 텍스트 조건부 생성 모델이며, ControlNet[4]은 깊이맵·에지맵 등의 조건을 주입해 구조를 보존한다. IP-Adapter[5]는 참조 이미지를 경량 모듈로 인코딩하여 Stable Diffusion에 전달, 텍스트와 결합된 정교한 합성을 가능하게 한다. 본 연구는 이 세 가지를 결합해 (1) 전처리 단계에서 아티팩트 제거와 색 보정을, (2) 증강 단계에서 주석 보존형 배경 합성을 수행함으로써 데이터 품질과 다양성을 동시에 확보한다.

2.2 제안기법

본 연구는 수중 객체 탐지를 위한 데이터 품질 개선과 자동화 증강을 위해 두 가지 핵심 기법을 제안한다. 2.2.1절에서는 원본 데이터의 저품질 문제(색 왜곡, 노이즈, 아티팩트)를 완화하기 위한 전처리 기법을 제안한다. 2.2.2절에서는 객체 위치를 보존한 상태에서 새로운 배경을 합성함으로써 대규모 학습 데이터를 확보할 수 있는 자동화 데이터 증강 파이프라인을 제안한다. **그림 1**은 본 연구의 데이터 전처리 및 증강 파이프라인을 나타낸 것이다.

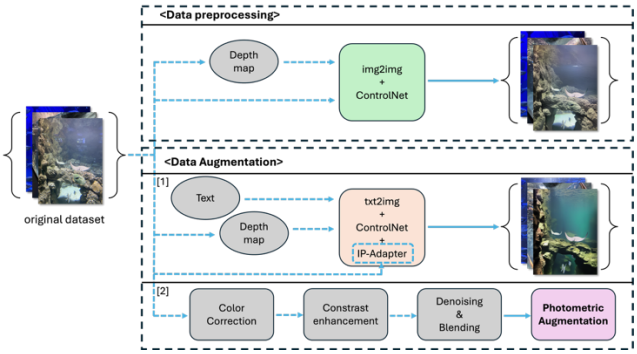

그림 1. 수중 객체 탐지를 위한 데이터 전처리 및 증강 파이프라인: 원본 데이터셋은 깊이 맵 기반 변환과 ControlNet[3]을 통해 전처리된다. 이후 두 가지 데이터 증강 방식이 적용된다: (1) 텍스트 조건과 IP-Adapter[5]를 활용한 diffusion 기반 증강, (2) 화이트밸런스, 대비 향상, 노이즈 억제 및 블렌딩을 통

* : These authors contributed equally.

한 외관(appearance) 다양성 확장을 위한 포토메트릭 증강.

2.2.1. 데이터 전처리 기법: Img2Img 기반 품질 개선

본 절에서는 Stable Diffusion(img2img)을 기반으로 ControlNet을 결합한 이미지 변환 기법을 적용하여, 저품질 수중 이미지의 색 왜곡과 노이즈를 완화하는 전처리 방법을 제안한다. 먼저 YOLO 라벨과 SAM[6] 모델을 활용해 객체를 분리하고, OpenCV/LaMa[7,8] 기반 inpainting을 통해 배경 이미지를 확보한다. 이후 이 배경 이미지를 초기 입력 I_{bg}으로 하여, 프롬프트 P와 깊이(depth) 조건 D를 주입해 조건부 확산 과정을 수행한다:

$$I' = \mathcal{F}_{\text{img2img}}(I_{bg}|P, D, \lambda)$$

여기서 λ는 원본 보존 정도를 조절하는 하이퍼파라미터이다. DDIM inversion[9] 기법을 응용하여 원본 이미지를 latent 공간으로 변환한 뒤, 노이즈 주입 및 조건 기반 복원을 수행함으로써 원본의 기하 구조는 유지하면서도 색 왜곡, 아티팩트, 저해상도 문제를 개선하였다. 이를 통해 보다 선명하고 안정적인 학습 데이터를 확보할 수 있었다. 제안한 전처리 기법의 효과는 **그림 2**에 나타나 있다. 전처리 결과, 원본의 구조는 유지되면서도 보다 선명하고 깨끗한 이미지를 얻을 수 있었으며, 촬영 과정에서 발생한 반사나 카메라 표면의 물방울과 같은 불필요한 아티팩트 또한 제거된 것을 확인할 수 있다.

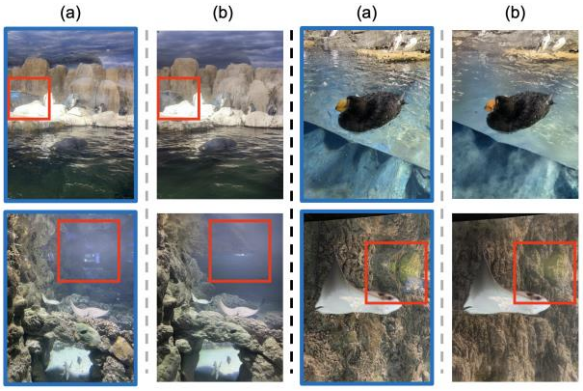

그림 2. 수중 이미지 전처리 결과 비교: (a) 원본 이미지, (b) Diffusion 기반 전처리를 통한 화질 개선 및 아티팩트가 제거된 이미지. 빨간 박스는 개선 효과가 두드러진 영역을 표시함.

2.2.2. 자동화 데이터 증강 파이프라인: Diffusion 및 Photometric Augmentation

본 절에서는 데이터 부족 문제를 해소하기 위해 Stable Diffusion(txt2img) + ControlNet + IP-Adapter 기반의 자동화 증강 파이프라인을 제안한다. 객체 마스크를 분리한 뒤, 텍스트 프롬프트 P, 깊이 맵 D, 그리고 IP-Adapter에 제공되는 참고 이미지(배경이미지) I_{ref}를 조건으로 하여 새로운 배경을 합성한다. 이 과정은 다음과 같이 표현할 수 있다:

$$I' = \mathcal{F}_{txt2img}(P, D, I_{ref}, \alpha)$$

여기서 α는 IP-Adapter의 스케일 파라미터로, 참조 이미지와 텍스트 프롬프트의 영향력을 조절한다. 이 과정은 랜덤 노이즈 $z_0 \sim \mathcal{N}(0, I)$에서 시작하므로, Img2Img와 달리 새로운 샘플을 생성하는 생성적(generative) 접근이다. 생성된 배경 위에 원래 객체를 삽입해, 객체·주석을 유지한 채 다양한 수중 환경을 합성할 수 있으며, **그림 3**에서 그 결과를 확인할 수 있다.

또한 본 연구에서는 외관 다양성 확장을 위해 포토메트릭 증강을 적용하였다. Diffusion 기반 증강이 새로운 환경 합성을 통한 도메인 확장이라면, 포토메트릭 증강은 색·대비·노이즈를 조정해 수중 영상 왜곡을 완화하고 신호 대 잡음비(SNR)와 색 충실도를 높인다. 두 방법은 상호 보완적이며, 2.3절에서 병행 효과를 확인할 수 있다.

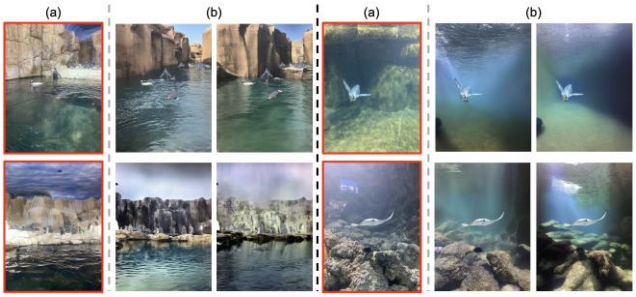

그림 3. 제안한 자동 증강 파이프라인 결과: (a) 원본 이미지(빨간 박스), (b) 조건부 확산 기반 합성 이미지.

2.3 실험결과

Kaggle Underwater Dataset(7 classes)으로 YOLOv8을 30 epoch 학습한 결과를 **표 1**에 제시한다. 데이터는 train 448장, validation 127장, test 63장으로 구성된다. 실험 조건은 Baseline(원본 데이터), 전처리(pre-proc), 포토메트릭 증강(aug1), Diffusion 기반 증강(aug2), 그리고 모든 기법을 결합한 Full로 설정하였다.

Ablation setting				Metrics	
Baseline	pre-proc	aug1	aug2	mAP@50 (↑)	mAP@50–95 (↑)
✓				0.7969	0.4527
✓	✓			0.8076 (+1.07%)	0.4773 (+2.46%)
✓	✓	✓		0.8387 (+4.18%)	0.5097 (+5.70%)
✓	✓		✓	0.8341 (+3.72%)	0.5116 (+5.89%)
✓	✓	✓	✓	**0.8612** (+6.43%)	**0.5303** (+7.76%)

표 1. 전처리(pre-proc) 및 증강 기법에 따른 mAP@50과 mAP@50–95 결과. aug1은 Photometric Augmentation, aug2는 Diffusion-based Augmentation을 나타내며, 각 기법의 조합 효과를 정량적으로 검증하였다.

Pre-proc은 1–3%의 개선을 보였으며, aug1과 aug2는 각각 독립적으로도 4–6%p 수준의 유의미한 성능 향상을 달성하였다. 특히 모든 기법을 결합한 Full은 Baseline 대비 6–8%p 향상된 성능을 보였다. 이는 포토메트릭 증강과 Diffusion 기반 증강이 개별적으로도 효과적이며, 결합 시 상호 보완적 시너지를 발휘하였음을 보여준다. 나아가, 제안한 방법이 데이터 부족과 품질 저하 문제를 동시에 완화할 수 있음을 실험적으로 입증한다. 또한, 단일 NVIDIA RTX 3090Ti GPU(24GB) 환경에서 이미지당 평균 처리 시간은 약 3초로, 사전 데이터 생성에는 충분히 효율적임을 확인하였다. 주요 하이퍼파라미터인 λ(원본 보존 정도)와 α(IP-Adapter 스케일)는 성능에 민감하게 작용하여, 극단적 설정에서는 아티팩트가 발생하였으나 $\lambda = 0.6$ ($img2img\_strength = 0.40$), $\alpha = 0.85$에서 안정적인 결과를 얻었다.

Ⅲ. 결론

본 연구는 수중 객체 탐지의 데이터 부족과 품질 저하 문제를 동시에 해결하기 위해 조건부 확산 기반 전처리와 자동 증강 파이프라인을 제안하였다. 색 왜곡·노이즈 보정과 다양한 배경 합성을 통해 추가적인 라벨링 없이 대규모 학습 데이터를 구축할 수 있었다. 실험 결과 YOLOv8 모델의 성능은 mAP@50 약 6%p, mAP@[0.50:0.95] 약 8%p 향상되었고, 각 기법은 독립적으로도 유효하며 결합 시 상호 보완적 효과를 보였다. 다만 test 63 장에 한정된 실험으로 일반화 검증에는 한계가 있으며, 향후 다양한 수중 데이터셋과 실제 환경 실험으로 보편성을 검증할 필요가 있다. 아울러 생성 과정에서 발생하는 아티팩트는 filtering 기반 후처리로 개선 가능해 데이터 품질을 높일 수

있으며, 본 접근법은 수중 탐지뿐 아니라 데이터 확보가 어려운 여러 비전 응용 분야로 확장될 수 있다.

참 고 문 헌

[1] Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLOv8 (Version 8.0.0) [Computer software]. GitHub. https://github.com/ultralytics/ultralytics

[2] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In Advances in Neural Information Processing Systems.

[3] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10684-10695).

[4] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 3813-3824).

[5] Ye, H., Zhang, J., Liu, S., Han, X., & Yang, W. (2023). IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models [Preprint]. arXiv.

[6] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything Model (SAM) [Computer software]. In Proceedings of the IEEE/CVF International Conference on Computer Vision.

[7] Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., & Lempitsky, V. (2021). Resolution-robust Large Mask Inpainting with Fourier Convolutions (LaMa) [Preprint]. arXiv.

[8] Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.

[9] Zeng, Y., Suganuma, M., & Okatani, T. (2025). Inverting the generation process of Denoising Diffusion Implicit Models: Empirical evaluation and a novel method. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 4616-4624). IEEE.

[10] Prytula, S. (2024). Aquarium data COTS dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/slavkoprytula/aquarium-data-cots