

Beyond Aesthetics: Quantifying Information Loss in Turbid Scenes

Vasiliki Ismiroglou^{1,2}, Stefan H. Bengtson^{1,2}, Tasos Benos¹, Thomas B. Moeslund^{1,2}, and Malte Pedersen^{1,2}

¹ Visual Analysis and Perception Laboratory, Aalborg University, Denmark

² Pioneer Centre for Artificial Intelligence, Denmark

Abstract. Visibility in underwater environments degrades rapidly under turbid conditions, yet the effects on computer-vision models remain unclear. This issue is compounded by reliance on synthetic turbidity datasets, which may misrepresent real-world information loss. To address this gap, we introduce the Turbid Underwater Baseline (TUB) dataset, comprising 1,320 images captured under extreme turbidity and over 16,000 high-confidence ground-truth segmentation masks. We additionally propose PCD, a metric derived from phase congruency maps that is invariant to contrast and aims to capture the loss of structural information in real turbidity. We show that PCD correlates strongly with the performance of instance segmentation models on both real and synthetic turbid images, whereas common metrics in the field show weak to no correlation at all. The dataset and relevant code can be found on the project page: <https://vap.aau.dk/pcd>

Keywords: Underwater · Turbidity · Metric · Phase congruency · Synthetic data

1 Introduction

Vision systems are increasingly used for critical infrastructure monitoring, biodiversity estimation, and navigation in diverse underwater environments [10, 21, 39, 42, 49]. These tasks often involve complex and low-visibility scenes, where light attenuation and scattering may dramatically reduce the perceptual quality of images. In such conditions, the performance of downstream computer vision models may decline drastically, which is typically attributed to perceptual degradation [16, 54, 56]. However, the reasoning for and level of the performance drop is not fully understood, as there currently exist no thoroughly annotated real-world image datasets covering the full range of underwater turbid conditions.

Turbidity refers to the cloudiness of water due to suspended particles and is a complex optical phenomenon with distinct visual effects. As light travels from the environment to the camera through the turbid medium, it undergoes absorption and scattering [36]. Absorption is strongly wavelength-dependent, leading to darker, spectrally distorted images [1, 47]. Scattering is a more complex process in which particles redirect light at angles determined by their size and shape.

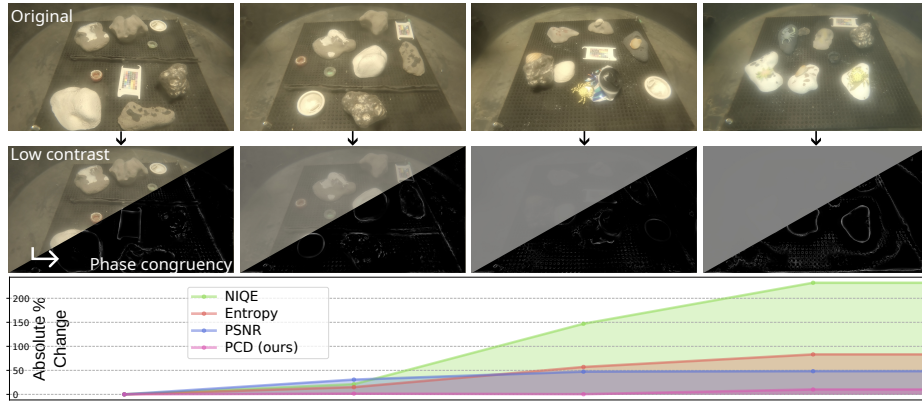


Fig. 1: Four clear images from the TUB dataset are shown with progressively reduced contrast. The proposed PCD metric indicates little change in information while alternative metrics show a substantial drop in image quality. The corresponding phase congruency maps, computed from the low contrast images, highlight that a considerable amount of structural information is retained.

As a result, peripheral light enters the camera in the form of backscatter, and light originating from the scene is either deflected away or spread across a wider area of the sensor [19, 33]. Because turbid mediums are rarely homogeneous, these effects are further convoluted due to the varying local density of sediment. The combined effect is a characteristic visual degradation: color distortion from selective absorption, reduced contrast due to the milky backscattering veil, and in highly scattering conditions *structural degradations*, in the form of blurred or otherwise corrupted object boundaries.

Owing to the scarcity of available turbid data, a substantial body of recent research has shifted toward the generation of synthetic images by modeling this light-medium interaction [5, 8, 30, 51]. Nevertheless, such works have predominantly concentrated on perceptual image enhancement applications, replicating only a subset of common visual degradations such as color shifts and contrast reduction. Without a framework to quantify the information loss that occurs in high-scattering environments, it is an open question whether these synthetic effects share any meaningful similarities to the structural degradations of real turbidity. If they do not, synthetic data may contribute little to training detection models and may even introduce harmful biases. To address these problems, we collect a new dataset designed to investigate the impact of turbidity on downstream computer vision tasks and propose a novel metric that can quantify this effect on real and synthetic data.

We present Phase-Congruency Delentropy (PCD), a metric of delentropy [24] using phase-congruency maps [23]. Phase congruency is invariant to color shifts and contrast loss, and instead captures the structural information of the image, an aspect largely overlooked in existing work on underwater image quality evaluation. Fig. 1 highlights how loss of contrast can result in a decrease in

perceptual quality while underlying structural information remains. In order to validate the metric, we introduce the first, *non-synthetic*, instance segmentation dataset featuring varying levels of turbidity, ranging from clear water to fully opaque conditions. We analyze how instance segmentation models behave under real turbidity and under commonly used synthetic degradations, demonstrating that PCD exhibits a strong correlation with downstream model performance. To summarize, our contributions are:

- The Turbid Underwater Baseline (TUB) dataset, with images systematically captured from multiple viewpoints and under varying levels of turbidity. It is the first publicly available dataset offering high-confidence ground-truth annotations in extremely turbid conditions, making it *suitable for data-driven tasks*.
- PCD: A novel metric for model-centric assessment of turbid scenes, demonstrating *superior correlation* with model performance, across both real and synthetic turbidity conditions compared with existing alternatives.
- We report baseline results for state-of-the-art instance segmentation models trained and evaluated across varying levels of real turbidity. They represent the first results to directly assess how turbidity affects learning-based models.
- We demonstrate that poor perceptual quality does not equate to poor task utility, whereas a reduction in structural information may.

2 Related Work

Datasets with Real Turbidity Publicly available datasets that capture a broad spectrum of real turbidity conditions remain limited and are often curated for image enhancement tasks [17, 25, 29, 48]. They therefore lack the annotations required to evaluate other downstream tasks, such as object detection or semantic segmentation. More critically, these datasets contain few, if any, examples of severe turbidity. The TURBID dataset [6] offers a wider range of induced turbidity levels, yet it encompasses only a small number of scenes, restricting its applicability for data-driven tasks as it contains less than 100 images in total.

When examining detection-centric datasets, the RUOD [9], UDD [27], and RUIE [29] datasets contain scenes that are commonly regarded as turbid. However, although the images exhibit strong color distortions, the targets themselves generally remain clear and well-defined. FishInTurbidWaters [20] includes fish imagery under more severe turbidity, but its annotations are sparse, providing only binary presence-absence labels. The Brackish dataset [42] offers bounding-box annotations in high-turbidity conditions, yet the range of environmental variability is relatively narrow. A key limitation of these datasets is that they do not record information about the environmental conditions, making it difficult to relate model performance to turbidity.

Synthetic Turbidity Generation In response to the scarcity of varied real-world data, synthetic turbidity generation has become a common strategy. Synthetic

turbid datasets generally fall into two categories: learning-based generative models [7,26,48,51], and physical-model-based methods that apply numerical approximations of a light scattering model directly to 2D images [1,5,15,18,30,50]. The latter, commonly termed as the underwater image formation model, is particularly widespread due to its computational simplicity and is often used to create training data for underwater image restoration.

The realism of the synthesized data is typically assessed indirectly, by using them to train image enhancement models and evaluating their ability to perceptually improve real images. Because current enhancement benchmarks rarely include extreme turbidity, synthetic pipelines are not optimized to reproduce such conditions despite their theoretical flexibility. As a result, characteristics unique to high turbidity environments, such as intense blur and heavy occlusions, are either absent from available synthetic data or not evaluated appropriately.

Image-Quality Metrics Image-quality metrics aim to approximate perceived quality and are often used to assess underwater images. These metrics can broadly be categorized according to the availability of reference data: full-reference (FR) metrics, where a degraded image is compared to a corresponding high-quality reference, and no-reference (NR) metrics, which attempt to estimate image quality without any ground truth.

Two of the most common full-reference metrics used in the underwater domain [2,17,25,44] are PSNR and SSIM [14]. While PSNR is strictly focused on quantifying noise, it is often used as a way to gauge the perceived quality of images. The obvious limitation of such metrics is that it is incredibly challenging to acquire true reference images underwater. In most existing cases, the reference and degraded images are artificially altered versions of each other.

Common NR metrics specifically designed for the underwater domain are UCIQE [55] and UIQM [41]. Both metrics are based on coefficients finetuned on specific data and may diverge if applied in drastically different conditions. BRISQUE [34], DIIVINE [37], and BLINDS [45] are examples of general-purpose NR metrics, which have been developed by training on databases of human-rated distortions. NIQE [35] was proposed as truly blind as it requires no such labels.

Image-Complexity Metrics Image-complexity metrics attempt to quantify the complexity or information in an image and do not necessarily consider its perceived quality. Entropy and Delentropy [24] are both information-centric NR measures that have been shown to correlate with model performance on common benchmarks [43]. Another example is VIF [46], which is a full-reference information-centric metric based on wavelet decomposition, and SSEQ [28] that extends similar principles to a no-reference, learning-based framework.

Prior work has shown that improving the perceived visual quality of underwater images does not translate into better detector performance [32,52]. This suggests that existing metrics do not adequately capture the information content that matters for data-driven learning in turbid conditions. As a result, these met-

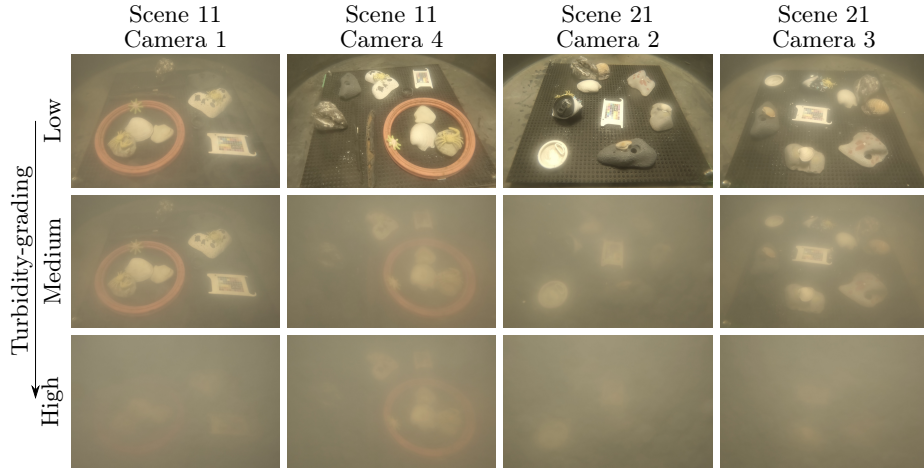


Fig. 2: Example of different scenes from the TUB dataset with different levels of turbidity (grouped by NTU) and from different camera-viewpoints.

rics are also likely unsuitable for evaluating whether synthetic turbidity effects reproduce the information loss observed in real environments.

These findings point to the need for a dedicated evaluation measure for turbid scenes. Developing such a measure requires a dataset that contains real images with varying levels of turbidity, which would allow the degradation process to be characterized and modeled. To address this gap, we collect the TUB dataset which we present in Sec. 3. The TUB dataset facilitates the evaluation of the novel PCD metric introduced in Sec. 4.

3 The Turbid Underwater (TUB) Dataset

We introduce the Turbid Underwater Baseline (TUB) dataset to address the gap in available datasets focusing on turbid underwater imagery. TUB is intended to support a wide range of research tasks involving high-turbidity environments, including but not limited to underwater image enhancement and restoration with explicit ground truth, object detection and segmentation under severe visibility degradation, and multi-view reconstruction. Examples of different scenes, viewpoints, and levels of turbidity in the dataset are shown in Fig. 2.

When designing a dataset to meaningfully evaluate the impact of turbidity on vision models, other sources of uncertainty must be minimized. Cameras and objects should remain fixed across turbidity levels to ensure that comparisons reflect only the effects of degradation. The turbidity itself must be both consistent and quantitatively measured to establish a reliable ground truth. Finally, building a dataset suitable for data-driven tasks requires not only sufficient image volume but also enough scene diversity to support robust learning.

Setup The data-collection setup for TUB was kept fully static, both in terms of camera placement and scene configuration. Four GoPro11 cameras were mounted around the perimeter of a cylindrical tank, and a LEGO baseplate was fixed at the bottom. Objects were attached to the baseplate using LEGO bricks and thus remained static throughout each turbidity sequence. This is the same approach as the BUCKET dataset [18].

Objects were selected randomly, including natural items such as rocks and man-made items such as trash. They were arranged in different positional configurations and object combinations, ensuring that no two scenes shared the same layout or orientations. These choices fulfilled the scene consistency requirement. The multi-camera setup enabled a substantial increase in image volume by capturing each scene from multiple viewpoints, naturally providing different occlusion levels without additional collection effort. This allowed more scenes to be gathered within the same timeframe.

Protocol Turbidity was introduced by filling the tank with water and adding measured amounts of oat milk. A total of 10 levels of turbidity were induced for every scene, with the first being clear water. A nephelometer was used to obtain accurate turbidity readings from three water samples at every stage.

Because turbidity levels varied across experiments, even when oat-milk increments were measured accurately, the same nominal stage (e.g., “stage 3”) could still appear noticeably different between scenes. Therefore, nephelometer readings were used to regroup the images into three turbidity categories: low (0–10 NTU), medium (10–20 NTU), and high (20+ NTU). This grouping serves two purposes. First, it produces a clear visual separation: objects are distinctly visible in the low group, partially degraded in the medium group, and heavily obscured in the high group. Second, the resulting distribution of images is relatively balanced across the three sets, making them suitable for training and evaluation experiments.

In total, the dataset contains 1,320 images, with 40 images per scene (4 viewpoints $\times$ 10 turbidity levels) across 33 unique scenes. This makes it the largest dataset of its kind, providing clear-image references while also capturing extreme turbidity conditions.

Annotations Manual annotations were created for all visible objects, resulting in over 16,000 instance-level segmentation masks. A single class label was used during annotation, with no distinction made between different object types. All scenes were annotated under clear (non-turbid) conditions using the images captured prior to adding oat milk. These annotations were then propagated to all corresponding turbid images, as the physical arrangement of each scene remained fixed across turbidity levels, as illustrated in Fig. 3.

This approach makes TUB the first underwater dataset to provide high confidence ground-truth labels even under severe turbidity, where human annotators would otherwise be unable to accurately delineate object boundaries.



Fig. 3: Annotation pipeline. Masks are delineated in the clear images and extended to all turbidity levels due to the static nature of the setup.

4 Delentropy of Phase Congruency Maps

While a vast range of image quality and complexity metrics exist and have been used in the underwater domain, the majority of them aim to align with human perception. However, deep-learning architectures have the theoretical capacity to extract information from images regardless of whether a human can perceive that same information.

Prior work has established that the principal structural information in an image resides in the phase component of its Fourier transform [40]. We hypothesize that this structural information is not entirely lost even when affected by attenuation and backscatter typical of underwater environments. We consider forms of discoloration and contrast distortion that still preserve shapes and region boundaries to represent a domain-shift phenomenon and therefore be largely recoverable by deep learning models. Conversely, when severe contrast loss or wavelength-dependent attenuation causes originally distinct neighboring regions to converge to identical pixel values, the underlying gradients may be irretrievably lost. Similar loss can occur when edges are heavily blurred or corrupted, causing boundaries to diffuse or disappear.

Delentropy already provides a framework to measure image complexity through the distribution of gradients. Originally rooted in the feature analysis of Haralick et al. [11] and later described by Larkin [24] as a natural entropy measure for images, delentropy is defined as

$$H(\nabla f) = H(f_x, f_y) = - \sum_{i=1}^J \sum_{j=1}^I p_{i,j} \log_2 p_{i,j} \quad (1)$$

where $p_{i,j}$ denotes the deldensity, the joint probability density of the gradient components f_x, f_y . In standard implementations, these gradients are estimated using fixed kernels. As a consequence, they are directly affected by contrast and illumination variations: regions with reduced contrast yield low gradient magnitudes, even when their underlying shapes remain discernible. Furthermore, the single-scale calculation of the gradients prevents the metric from capturing structural cues across different spatial extents.

To address these challenges, we investigate the use of phase congruency representations. Inspired by Morrone et al.'s [38] observation that salient features occur where local frequency components are maximally in phase, Kovess [23] pro-

posed phase congruency as a low-level image invariant demonstrating robustness to illumination and contrast variations.

Phase congruency maps produce values in the range $[0,1]$ and are computed across multiple scales. This multiscale formulation yields intensity variations that reflect edge sharpness, while still capturing broader or blurred boundaries. Visual examples of phase congruency images can be seen in Fig. 1.

Kovesi’s implementation employs wavelets derived from Gabor functions and includes mechanisms for noise suppression in the estimation process. While phase congruency can be computed in an arbitrary number of orientations for 2D images, we follow the derivation used in delentropy and compute it over two orthogonal orientations. This produces phase-congruency-based gradient maps. By replacing the kernel-derived gradients used in delentropy with these phase congruency maps, we introduce a new metric that we refer to as PCD.

5 Experiments

5.1 TUB Baseline

We first perform baseline experiments on the TUB dataset to evaluate model performance and behavior on real turbid images. We partition the data into splits at the scene level, keeping 8 scenes for validation and 25 for training. This ensures that while the same objects may appear in both splits, they never do so in the same spatial configuration or position. We further create subsets based on the measured turbidity values by dividing the images into low, medium, and high turbidity, as described earlier in Sec. 3.

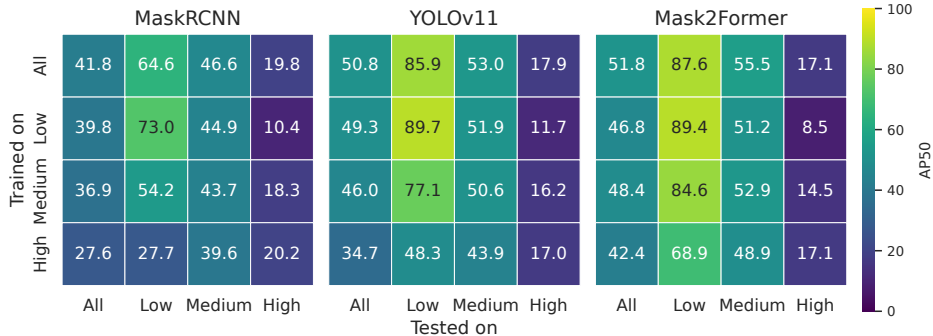
We focus on instance segmentation in the evaluation, as it is a common task within underwater computer vision. We train and evaluate three state-of-the-art models, MaskRCNN [12], YOLOv11 [22], and Mask2Former [4]. These collectively cover both major architecture types for vision, i.e., convolutional neural networks and transformers. Implementation details vary between architectures due to sensitivities to learning rate and memory limitations. At the chosen training iterations, all models had converged with minor variations in losses and validation metrics. Details can be found in Tab. 1. The results in the form of AP50 are shown in Fig. 4, highlighting the relationship between model performance and measured turbidity. Furthermore, we show some qualitative examples of predicted masks in Fig. 5.

We found the behavior across models to be very consistent despite the variations in absolute performance. Our results indicate that an increase in turbidity affects predictive capabilities even when there is no domain shift to consider. Nevertheless, all models saw improvements in the *medium* and *high* turbidity settings when the respective data were available during training. Extremely turbid data appear to be introducing ambiguity to the model, with performance dropping for the clean images when included. This does not affect all models equally.

³ MaskRCNN and Mask2Former were trained using the detectron2 [53] framework, where 6000 steps are used instead of epochs.

Table 1: Training configuration of the different instance segmentation models.

Model	Arch.	Epochs ³	LR	Batch
MaskRCNN	ResNet101-FPN [13]	48	10^{-5}	8
YOLOv11	YOLO11m-seg	50	10^{-3}	32
Mask2Former	SwinB [31]	48	10^{-4}	8

**Fig. 4:** The performance reported as AP50 for MaskRCNN, YOLOv11, and Mask2Former when trained and tested on low, medium, and high turbidity images from the TUB dataset.

5.2 PCD Evaluation

We additionally aim to evaluate how well existing image- quality and complexity metrics, including our proposed PCD, correlate with performance in downstream tasks when using turbid data. For this task we include both real images and physical-model-based synthetic data. The intention is not to perform a systematic assessment of the synthetic data itself, but to examine how simplified representations of real degradation phenomena influence the behavior of these metrics.

Using the same scene-based split described in Sec. 5.1, we take the lowest turbidity images and generate ten synthetic variants per image through two different 2D, image-based, synthetic degradation methods. The first applies the underwater image-formation model [1, 5], treating turbidity primarily as distance-dependent contrast loss and will be named *Synth*₁ for ease of reference. The second model, *Synth*₂ [18], extends this approach by incorporating blur and spatially varying inhomogeneity noise.

Image formation model approaches require scene geometry and attenuation coefficients to produce images. We employ ZoeDepth [3] as a monocular depth estimator and randomly rescale its outputs between 0.2m and 4m. We also sample attenuation coefficients from a distribution derived from the Jerlov water types [47]. This setup reflects typical scenarios where scene geometry and the exact model parameters are unknown. The synthetic images retain the same

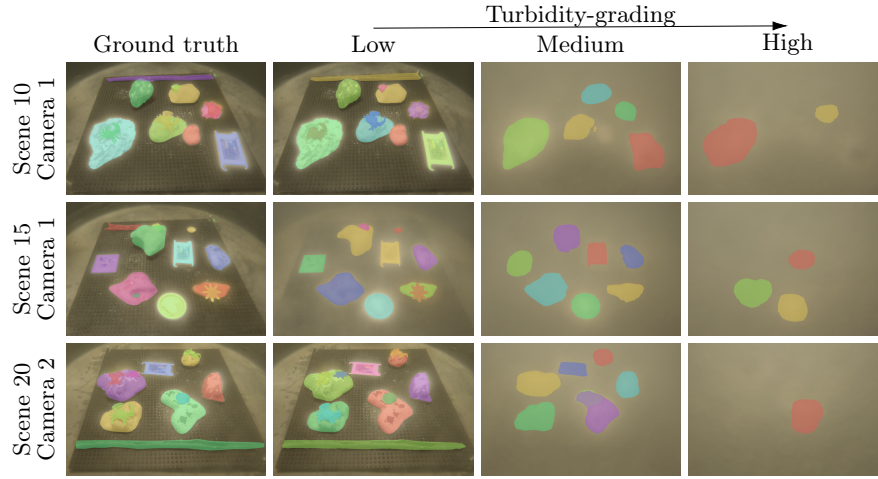


Fig. 5: Example of predicted masks from the Mask2Former model trained on the TUB dataset. All images are from the validation split. Colors are for visualization only.

mask annotations as their real counterparts, and both sets are merged to form an expanded dataset. In total, this triples the volume of the images.

Similar to the turbidity baseline, we train MaskRCNN, YOLOv11, and Mask2Former on the combined dataset and report AP50 separately per image. We also compute a broad set of quality and information-content metrics. Shannon entropy and Delentropy are included because they directly relate to the proposed PCD measure as indicators of information and structural complexity. In addition, we compute three no-reference quality metrics: NIQE, which is designed for general natural-image statistics, and UIQM and UCIQE, which are tailored to underwater imagery. Finally, we measure SSIM and PSNR using the lowest-turbidity images as a reference to quantify structural and photometric deviations. We do not fine-tune the statistics or coefficients of any of the metrics on our data.

The correlation of different metrics with model performance is summarized in Fig. 6 while the full scatterplots are shown in Fig. 7. We find that many of the metrics exhibit weak correlation even on real images. Turbid scenes often collapse into narrow metric ranges despite clear variation in object visibility as reflected by model performance. We also observe a consistent pattern across models: MaskRCNN shows the lowest correlations with all metrics, followed by YOLOv11 and Mask2Former, a hierarchy that aligns with their predictive capabilities.

When evaluating image quality on synthetic data, all established metrics fail to reflect the actual degree of degradation, as shown in Fig. 7. A recurring trend is that the synthetic images, particularly the ones of *Synth*₁, are scored as substantially lower quality than real turbid images, despite the models maintaining high-precision predictions on many of them. On the contrary, PCD produces

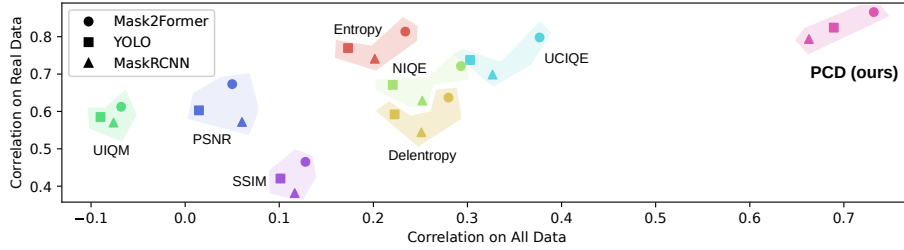


Fig. 6: Metric correlation with model performance. The area highlights do not represent any measurable quantity, but are simply used to assist visualization.

higher scores, similar to the ones of clear images, for the majority of $Synth_1$. While this results in some outliers, there is also a meaningful portion of overlap with the ground-truth distribution.

A visual example to explain this behavior can be seen in Figs. 1 and 8. When contrast is reduced without any blur or noise, gradients remain sharp. The majority of metrics already consider that a drop in quality, but model performance is not affected drastically. After a point, objects further from the camera disappear completely. Gradients are still sharp and indicative of a clear image, but detector performance drops as they are evaluated on objects that are fully occluded rather than degraded.

Overall, PCD shows the highest correlation with model performance, indicating that gradient degradation is a better indicator of image quality in high turbidity settings than contrast, discolorations, or pixel statistics. This is further supported by the differences in the synthetic data when gradient-degrading effects are considered.

Phase-congruency hyperparameters PCD does not rely on training data or statistical priors, but its phase-congruency computation includes hyperparameters that can be tuned. Kovési [23] noted that the impact of the hyperparameters is limited and illustrated his point with a few qualitative examples. Nevertheless, we believe that in images dominated by noise and diffuse gradients, such as those typical of turbid environments, the wavelet characteristics over which phase congruency is calculated are critical. Using only a few scales emphasizes high-frequency structure, whereas increasing the number of scales progressively incorporates lower-frequency information and diminishes the relative weight of fine-scale content.

To assess this sensitivity, we analyze PCD’s correlation with Mask2Former across a range of image-rescaling factors and wavelet-scale configurations. This allows us to evaluate how robust the metric remains under different parameter choices. The results are presented in Fig. 9. For the experiments reported in the main text, we downsampled the images by a factor of 0.5 and used seven wavelet scales with a minimum wavelength of 3 and a multiplication factor of

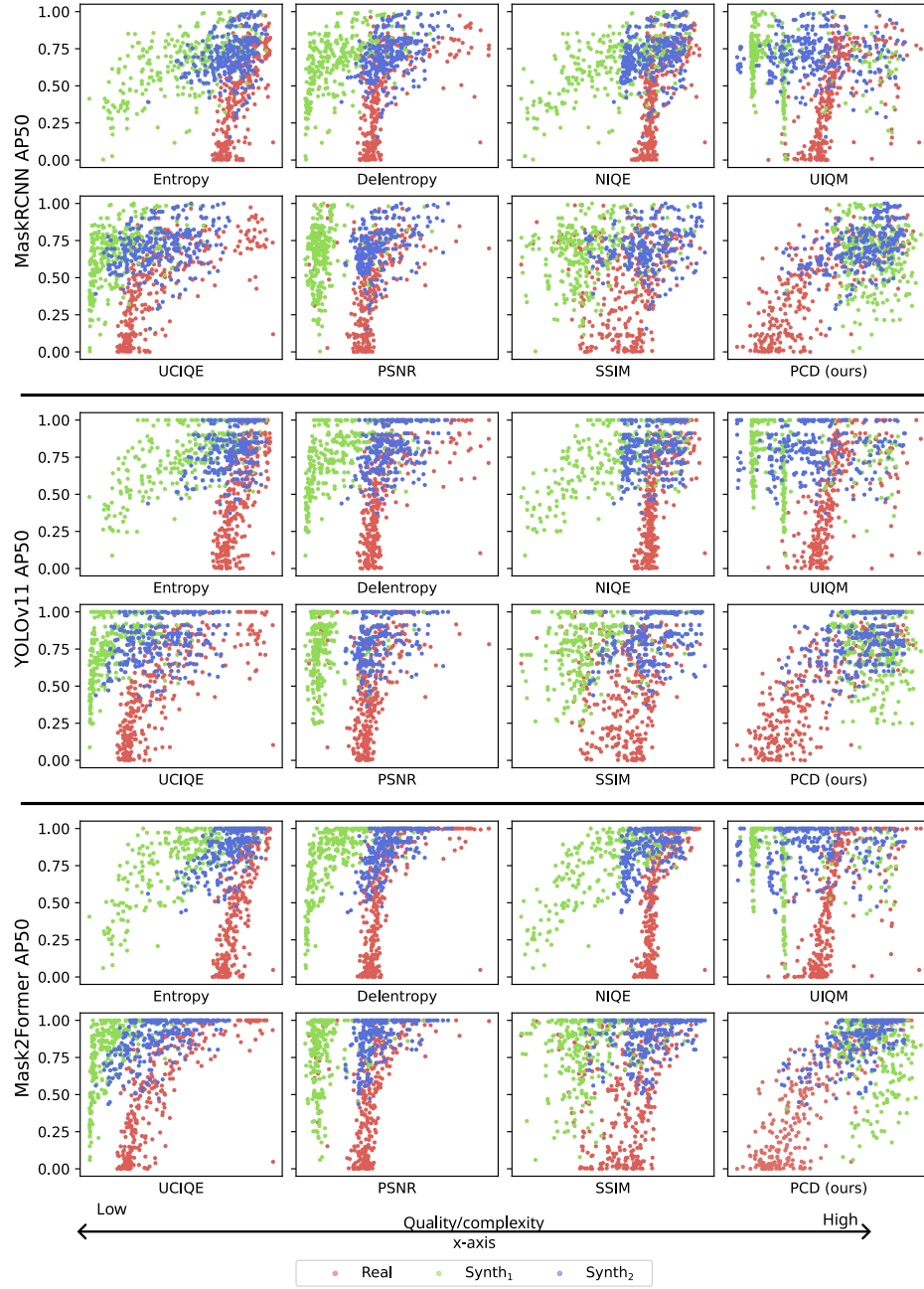


Fig. 7: Scatterplots highlighting correlation between model performance and common image quality metrics for real and synthetic data. The models listed from top to bottom are MaskRCNN, YOLOv11, and Mask2Former. NIQE is inverted for visual consistency.

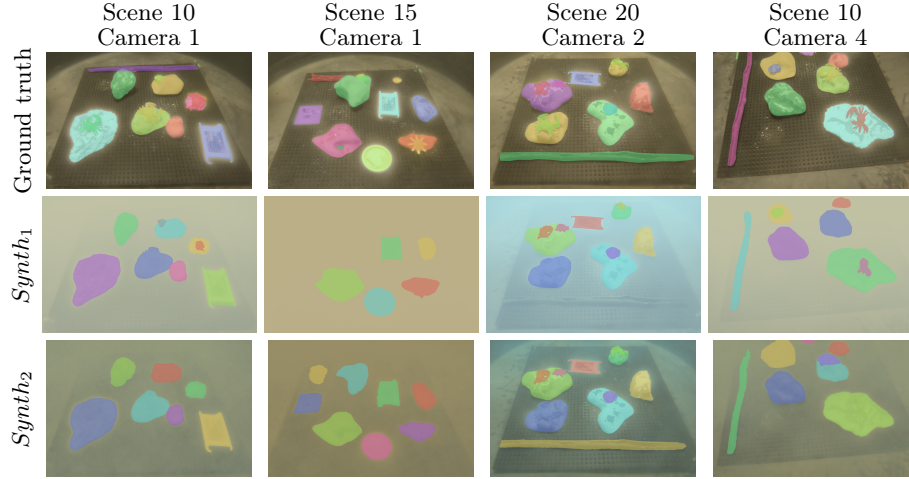


Fig. 8: Example of predicted masks on synthetic images. First row: ground truth masks on non-turbid images from the TUB dataset. Second and third row: predicted masks on images with synthetic turbidity generated using *Synth₁* and *Synth₂*, respectively. Predictions are from Mask2Former, and all images are from the validation split. Colors are for visualization only.

2.1. However, our analysis indicates that six wavelet scales with a minimum wavelength of 2 provide more stable behavior overall.

6 Discussion

Baseline turbidity experiments show that clear images are essential, as models trained solely on turbid data can exhibit reduced performance even in the absence of a domain gap. This likely reflects the role of clear images in enabling stronger feature representations, which have the capacity to generalize to novel viewpoints and conditions. Training on turbid data facilitates this generalization to degraded environments, often at the cost of reduced performance in clear im-

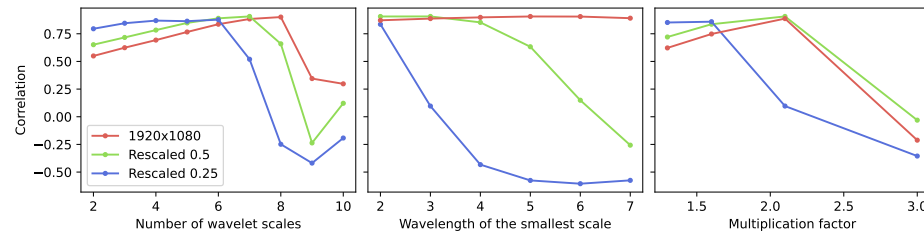


Fig. 9: Variation of PCD correlation to Mask2Former AP50 as the wavelet characteristic over which phase congruency is calculated change.

ages. This trade-off underscores the need for strategies that introduce degraded imagery into training pipelines without compromising performance across the full range of operating conditions.

The results also expose a practical challenge: incorporating highly turbid images into training becomes difficult when the turbidity level prevents reliable human annotation. While synthetic data can help tackle this problem, our findings show that commonly used synthetic degradations are not interpreted by existing image-quality metrics in the same way as real turbidity. Through our experiments, we demonstrate that structural degradations are more critical to detection models than perceptual ones, making PCD a better-suited metric at quantifying the effects of both real and synthetic turbidity on downstream tasks.

Caveats Non-reference metrics, particularly those intended to approximate information content, are inevitably influenced by the underlying scene structure, independent of degradation. This makes comparisons unreliable when images differ substantially in inherent detail, such as between a featureless, clear-water region and a highly textured scene under strong turbidity. In our case, the dataset was collected under controlled conditions, and all scenes contain comparable levels of structural detail. This consistency makes relative comparisons between metrics and degradation methods meaningful.

The dataset design also enables the isolation of error sources. Because the model’s predictions on the corresponding clear images are available, they serve as a reference to distinguish errors caused specifically by turbidity from those arising due to other factors, such as out-of-distribution content or occlusions. Without these conditions, evaluating such correlations would be far less straightforward, as it would be difficult to attribute failure cases to degradation rather than unrelated scene variations.

7 Conclusion

In this work, we introduce and evaluate PCD, a novel metric that focuses on structural degradation caused by turbidity rather than perceptual changes. Our experiments demonstrate that PCD exhibits the strongest correlation with performance on instance segmentation models across both real and synthetic turbidity when compared with existing alternatives. We thus highlight its potential for assessing turbid data information content as well as evaluating the realism of synthetic turbidity generation. We additionally present TUB, the first publicly-available instance segmentation dataset designed for extreme turbidity conditions and suitable for data-driven learning tasks such as the PCD experiments. This dataset addresses a major gap in the existing literature and provides an essential resource for advancing vision research in turbid underwater environments.

Acknowledgements

This work was funded by the Pioneer Centre for AI, DNRF grant number P1. We further gratefully acknowledge DTU Aqua for providing access to their nephelometer for TUB measurements.

References

1. Akkaynak, D., Treibitz, T.: A Revised Underwater Image Formation Model. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6723–6732 (Jun 2018). <https://doi.org/10.1109/CVPR.2018.00703>, <https://ieeexplore.ieee.org/document/8578801>
2. Anwar, S., Li, C., Porikli, F.: Deep Underwater Image Enhancement (Jul 2018). <https://doi.org/10.48550/arXiv.1807.03528>, <http://arxiv.org/abs/1807.03528>, arXiv:1807.03528 [cs]
3. Bhat, S.F., Birkel, R., Wofk, D., Wonka, P., Müller, M.: ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth (Feb 2023). <https://doi.org/10.48550/arXiv.2302.12288>, <http://arxiv.org/abs/2302.12288>, arXiv:2302.12288 [cs]
4. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention Mask Transformer for Universal Image Segmentation (Jun 2022). <https://doi.org/10.48550/arXiv.2112.01527>, <http://arxiv.org/abs/2112.01527>, arXiv:2112.01527 [cs]
5. Desai, C., Benur, S., Patil, U., Mudanagudi, U.: RSUIGM: Realistic Synthetic Underwater Image Generation with Image Formation Model. ACM Trans. Multimedia Comput. Commun. Appl. **21**(1), 10:1–10:22 (Dec 2024). <https://doi.org/10.1145/3656473>, <https://doi.org/10.1145/3656473>
6. Duarte, A., Codevilla, F., Gaya, J.D.O., Botelho, S.S.C.: A dataset to evaluate underwater image restoration methods. In: OCEANS 2016 - Shanghai. pp. 1–6 (Apr 2016). <https://doi.org/10.1109/OCEANSAP.2016.7485524>, <https://ieeexplore.ieee.org/abstract/document/7485524>
7. Fabbri, C., Islam, M.J., Sattar, J.: Enhancing Underwater Imagery Using Generative Adversarial Networks. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 7159–7165 (May 2018). <https://doi.org/10.1109/ICRA.2018.8460552>, <https://ieeexplore.ieee.org/document/8460552>
8. Fathy, M.E., Mohamed, S.A., Awad, M.I., Munim, H.E.A.E.: SubmergeStyleGAN: Synthetic Underwater Data Generation with Style Transfer for Domain Adaptation. In: 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 546–553 (Nov 2023). <https://doi.org/10.1109/DICTA60407.2023.00081>, <https://ieeexplore.ieee.org/document/10410962/>
9. Fu, C., Liu, R., Fan, X., Chen, P., Fu, H., Yuan, W., Zhu, M., Luo, Z.: Rethinking general underwater object detection: Datasets, challenges, and solutions. Neurocomputing **517**, 243–256 (Jan 2023). <https://doi.org/10.1016/j.neucom.2022.10.039>, <https://www.sciencedirect.com/science/article/pii/S0925231222013169>
10. González-Sabbagh, S.P., Robles-Kelly, A.: A Survey on Underwater Computer Vision. ACM Comput. Surv. **55**(13s), 268:1–268:39 (Jul 2023). <https://doi.org/10.1145/3578516>, <https://dl.acm.org/doi/10.1145/3578516>

11. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3**(6), 610–621 (Nov 1973). <https://doi.org/10.1109/TSMC.1973.4309314>, <https://ieeexplore.ieee.org/document/4309314>
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN (Jan 2018). <https://doi.org/10.48550/arXiv.1703.06870>, <http://arxiv.org/abs/1703.06870>, arXiv:1703.06870 [cs]
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (Dec 2015). <https://doi.org/10.48550/arXiv.1512.03385>, <http://arxiv.org/abs/1512.03385>, arXiv:1512.03385 [cs]
14. Horé, A., Ziou, D.: Image Quality Metrics: PSNR vs. SSIM. In: 2010 20th International Conference on Pattern Recognition. pp. 2366–2369 (Aug 2010). <https://doi.org/10.1109/ICPR.2010.579>, <https://ieeexplore-ieee-org.zorac.aub.aau.dk/document/5596999>, iSSN: 1051-4651
15. Hou, G., Zhao, X., Pan, Z., Yang, H., Tan, L., Li, J.: Benchmarking Underwater Image Enhancement and Restoration, and Beyond. *IEEE Access* **8**, 122078–122091 (2020). <https://doi.org/10.1109/ACCESS.2020.3006359>, <https://ieeexplore.ieee.org/document/9130676>, conference Name: IEEE Access
16. Hu, K., Weng, C., Zhang, Y., Jin, J., Xia, Q.: An Overview of Underwater Vision Enhancement: From Traditional Methods to Recent Deep Learning. *Journal of Marine Science and Engineering* **10**(2) (Feb 2022). <https://doi.org/10.3390/jmse10020241>, <https://www.mdpi.com/2077-1312/10/2/241>
17. Islam, M.J., Xia, Y., Sattar, J.: Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robotics and Automation Letters* **5**(2), 3227–3234 (Apr 2020). <https://doi.org/10.1109/LRA.2020.2974710>, <https://ieeexplore.ieee.org/document/9001231>
18. Ismiroglou, V., Pedersen, M., Bengtson, S.H., Aakerberg, A., Moeslund, T.B.: Sealing Through Scattered Rays: Revisiting the Image Formation Model for Realistic Underwater Image Generation. In: In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2025)
19. Jaffe, J.: Computer modeling and the design of optimal underwater imaging systems. *IEEE Journal of Oceanic Engineering* **15**(2), 101–111 (Apr 1990). <https://doi.org/10.1109/48.50695>, <https://ieeexplore-ieee-org.zorac.aub.aau.dk/document/50695>
20. Jahanbakht, M., Rahimi Azghadi, M., Waltham, N.J.: Semi-supervised and weakly-supervised deep neural networks and dataset for fish detection in turbid underwater videos. *Ecological Informatics* **78**, 102303 (Dec 2023). <https://doi.org/10.1016/j.ecoinf.2023.102303>, <https://www.sciencedirect.com/science/article/pii/S1574954123003321>
21. Jiao, P., Ye, X., Zhang, C., Li, W., Wang, H.: Vision-based real-time marine and offshore structural health monitoring system using underwater robots. *Computer-Aided Civil and Infrastructure Engineering* **39**(2), 281–299 (2024). <https://doi.org/10.1111/mice.12993>, <https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12993>, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/mice.12993>
22. Jocher, G., Qiu, J., Chaurasia, A.: Ultralytics YOLO (Jan 2023), <https://github.com/ultralytics/ultralytics>, original-date: 2022-09-11T16:39:45Z
23. Kovési, P.: Phase congruency: A low-level image invariant. *Psychological Research* **64**(2), 136–148 (Dec 2000). <https://doi.org/10.1007/s004260000024>, <https://doi.org/10.1007/s004260000024>

24. Larkin, K.G.: Reflections on Shannon Information: In search of a natural information-entropy for images (Sep 2016). <https://doi.org/10.48550/arXiv.1609.01117>, <http://arxiv.org/abs/1609.01117>, arXiv:1609.01117 [cs]
25. Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., Tao, D.: An Underwater Image Enhancement Benchmark Dataset and Beyond. *IEEE Transactions on Image Processing* **29**, 4376–4389 (2020). <https://doi.org/10.1109/TIP.2019.2955241>, <https://ieeexplore.ieee.org/document/8917818>
26. Li, N., Zheng, Z., Zhang, S., Yu, Z., Zheng, H., Zheng, B.: The Synthesis of Unpaired Underwater Images Using a Multistyle Generative Adversarial Network. *IEEE Access* **6**, 54241–54257 (2018). <https://doi.org/10.1109/ACCESS.2018.2870854>, <https://ieeexplore.ieee.org/abstract/document/8467323>
27. Liu, C., Wang, Z., Wang, S., Tang, T., Tao, Y., Yang, C., Li, H., Liu, X., Fan, X.: A New Dataset, Poisson GAN and AquaNet for Underwater Object Grabbing. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(5), 2831–2844 (May 2022). <https://doi.org/10.1109/TCSVT.2021.3100059>, <http://arxiv.org/abs/2003.01446>, arXiv:2003.01446 [cs]
28. Liu, L., Liu, B., Huang, H., Bovik, A.C.: No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication* **29**(8), 856–863 (Sep 2014). <https://doi.org/10.1016/j.image.2014.06.006>, <http://pure.bit.edu.cn/en/publications/no-reference-image-quality-assessment-based-on-spatial-and-spectr/>
29. Liu, R., Fan, X., Zhu, M., Hou, M., Luo, Z.: Real-world Underwater Enhancement: Challenges, Benchmarks, and Solutions (Mar 2019). <https://doi.org/10.48550/arXiv.1901.05320>, <http://arxiv.org/abs/1901.05320>, arXiv:1901.05320 [cs]
30. Liu, Y., Xu, H., Zhang, B., Sun, K., Yang, J., Li, B., Li, C., Quan, X.: Model-Based Underwater Image Simulation and Learning-Based Underwater Image Enhancement Method. *Information* **13**(4), 187 (Apr 2022). <https://doi.org/10.3390/info13040187>, <https://www.mdpi.com/2078-2489/13/4/187>
31. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows (Aug 2021). <https://doi.org/10.48550/arXiv.2103.14030>, <http://arxiv.org/abs/2103.14030>, arXiv:2103.14030 [cs]
32. Lucas, E., Awad, A., Gaglio, A., Moradi, S., Saleem, A., Havens, T., Galloway, A., Paheding, S.: Underwater Image Enhancement and Object Detection: Are Poor Object Detection Results On Enhanced Images Due to Missing Human Labels? In: *In Proceedings of the Winter Conference on Applications of Computer Vision Workshops* (2025)
33. McGlamery, B.L.: A Computer Model For Underwater Camera Systems. In: *Ocean Optics VI*. vol. 0208, pp. 221–231. SPIE (Mar 1980). <https://doi.org/10.1117/12.958279>, <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/0208/0000/A-Computer-Model-For-Underwater-Camera-Systems/10.1117/12.958279.full>
34. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing* **21**(12), 4695–4708 (Dec 2012). <https://doi.org/10.1109/TIP.2012.2214050>, <http://ieeexplore.ieee.org/document/6272356/>
35. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “Completely Blind” Image Quality Analyzer. *IEEE Signal Processing Letters* **20**(3), 209–212 (Mar 2013). <https://doi.org/10.1109/LSP.2012.2227726>, <http://ieeexplore.ieee.org/document/6353522/>

36. Mobley, C.: Light and Water: Radiative Transfer in Natural Waters. Academic (Jan 1994), journal Abbreviation: Academic Press Publication Title: Academic Press
37. Moorthy, A.K., Bovik, A.C.: Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. *IEEE Transactions on Image Processing* **20**(12), 3350–3364 (Dec 2011). <https://doi.org/10.1109/TIP.2011.2147325>, <https://ieeexplore.ieee.org/document/5756237/>
38. Morrone, M.C., Ross, J., Burr, D.C., Owens, R.: Mach bands are phase dependent. *Nature* **324**, 250–253 (Nov 1986). <https://doi.org/10.1038/324250a0>, <https://ui.adsabs.harvard.edu/abs/1986Natur.324..250M>, aDS Bibcode: 1986Natur.324..250M
39. Naveen, P.: Advancements in underwater imaging through machine learning: Techniques, challenges, and applications. *Multimedia Tools and Applications* **84**(22), 24839–24858 (Jul 2025). <https://doi.org/10.1007/s11042-024-20091-4>, <https://doi.org/10.1007/s11042-024-20091-4>
40. Oppenheim, A., Lim, J.: The importance of phase in signals. *Proceedings of the IEEE* **69**(5), 529–541 (May 1981). <https://doi.org/10.1109/PROC.1981.12022>, <https://ieeexplore.ieee.org/document/1456290>
41. Panetta, K., Gao, C., Agaian, S.: Human-Visual-System-Inspired Underwater Image Quality Measures. *IEEE Journal of Oceanic Engineering* **41**(3), 541–551 (Jul 2016). <https://doi.org/10.1109/JOE.2015.2469915>, <https://ieeexplore-ieee-org.zorac.aub.aau.dk/document/7305804>
42. Pedersen, M., Bruslund Haurum, J., Gade, R., Moeslund, T.B.: Detection of Marine Animals in a New Underwater Dataset with Varying Visibility. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. pp. 18–26 (2019), https://openaccess.thecvf.com/content_CVPRW_2019/html/AAMVEM/Pedersen_Detection_of_Marine_Animals_in_a_New_Underwater_Dataset_with_CVPRW_2019_paper.html
43. Rahane, A.A., Subramanian, A.: Measures of Complexity for Large Scale Image Datasets. In: *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*. pp. 282–287 (Feb 2020). <https://doi.org/10.1109/ICAIIIC48513.2020.9065274>, <http://arxiv.org/abs/2008.04431>, arXiv:2008.04431 [cs]
44. Raveendran, S., Patil, M.D., Birajdar, G.K.: Underwater image enhancement: a comprehensive review, recent trends, challenges and applications. *Artificial Intelligence Review* **54**(7), 5413–5467 (Oct 2021). <https://doi.org/10.1007/s10462-021-10025-z>, <https://doi.org/10.1007/s10462-021-10025-z>
45. Saad, M.A., Bovik, A.C., Charrier, C.: A DCT Statistics-Based Blind Image Quality Index. *IEEE Signal Processing Letters* **17**(6), 583–586 (Jun 2010). <https://doi.org/10.1109/LSP.2010.2045550>, <https://ieeexplore.ieee.org/document/5430991/>
46. Sheikh, H., Bovik, A.: Image information and visual quality. *IEEE Transactions on Image Processing* **15**(2), 430–444 (Feb 2006). <https://doi.org/10.1109/TIP.2005.859378>, <https://ieeexplore.ieee.org/document/1576816/>
47. Solonenko, M.G., Mobley, C.D.: Inherent optical properties of Jerlov water types. *Applied Optics* **54**(17), 5392–5401 (Jun 2015). <https://doi.org/10.1364/AO.54.005392>, <https://opg.optica.org/ao/abstract.cfm?uri=ao-54-17-5392>
48. Sun, W., Hou, S., Wu, G., Zhang, J., Zhang, Y., Jiang, H.: Turbid image tackling framework towards underwater concrete bridge detection based on distance control and deep learning. *Advanced Engineering Informatics* **62**, 102723 (Oct 2024). <https://doi.org/10.1016/j.aei.2024.102723>, <https://linkinghub.elsevier.com/retrieve/pii/S1474034624003719>

49. Álvarez Tuñón, O., Rodríguez, A., Jardón, A., Balaguer, C.: Underwater Robot Navigation for Maintenance and Inspection of Flooded Mine Shafts. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1482–1487 (Oct 2018). <https://doi.org/10.1109/IROS.2018.8594445>, <https://ieeexplore.ieee.org/document/8594445/>, iSSN: 2153-0866
50. Ueda, T., Yamada, K., Tanaka, Y.: Underwater Image Synthesis from RGB-D Images and its Application to Deep Underwater Image Restoration. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 2115–2119 (Sep 2019). <https://doi.org/10.1109/ICIP.2019.8803195>, <https://ieeexplore.ieee.org/document/8803195/>
51. Wang, N., Zhou, Y., Han, F., Zhu, H., Zheng, Y.: UWGAN: Underwater GAN for Real-world Underwater Color Restoration and Dehazing. ArXiv (Sep 2019), <https://www.semanticscholar.org/paper/ad8fbe003c65c15429ffa1a81fbd311f9bbae777>
52. Wang, Y., Guo, J., He, W., Gao, H., Yue, H., Zhang, Z., Li, C.: Is Underwater Image Enhancement All Object Detectors Need? IEEE Journal of Oceanic Engineering **49**(2), 606–621 (Apr 2024). <https://doi.org/10.1109/JOE.2023.3302888>, <https://ieeexplore.ieee.org/document/10376393/>
53. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. <https://github.com/facebookresearch/detectron2> (2019)
54. Xu, S., Zhang, M., Song, W., Mei, H., He, Q., Liotta, A.: A systematic review and analysis of deep learning-based underwater object detection. Neurocomputing **527**, 204–232 (Mar 2023). <https://doi.org/10.1016/j.neucom.2023.01.056>, <https://linkinghub.elsevier.com/retrieve/pii/S0925231223000656>
55. Yang, M., Sowmya, A.: An Underwater Color Image Quality Evaluation Metric. IEEE Transactions on Image Processing **24**(12), 6062–6071 (Dec 2015). <https://doi.org/10.1109/TIP.2015.2491020>, <https://ieeexplore.ieee.org/document/7300447>
56. Yuan, X., Guo, L., Luo, C., Zhou, X., Yu, C.: A Survey of Target Detection and Recognition Methods in Underwater Turbid Areas. Applied Sciences **12**(10) (May 2022). <https://doi.org/10.3390/app12104898>, <https://www.mdpi.com/2076-3417/12/10/4898>