Underwater Salient Object Detection Using Color Balance and Fusion Enhanced USOD10K Dataset and TC-USOD Model

1st Alluri Lakshman Narendra
Computer Science and Engineering
IIT Dharwad
Dharwad, India
220010002@iitdh.ac.in

2nd Attunuri Praneeth Reddy
Mechanical Engineering
IIT Dharwad
Dharwad, India
220030005@iitdh.ac.in

3rd Kapse Karthik
Electrical Engineering
IIT Dharwad
Dharwad, India
220020026@iitdh.ac.in

***Abstract*—While the underwater environment is very challenging for certain key object identification tasks, due to problems related to light absorption and color distortion. So one has to process the resulting frames from an underwater video something that is often suspected to be very unclear or blurry in nature. This makes it especially challenging for marine researchers whose work aims to study coral reefs or track the movements of marine animals. To eliminate this issue, we use the USOD10K dataset with preprocessing inspired by "Color Balance and Fusion for Underwater Image Enhancement". In short, we were able to boost the underwater images by correcting and cleaning color channels, which could retain them from being washed out. This improvement helped us better as marine creatures and objects, including in bad lighting or turbid water conditions. We then trained and tested a model to detect underwater objects with the enhanced images. We showed that our approach points to new possibilities: achieving higher accuracy than existing approaches and more visually pleasing results. Efforts to make discoveries under the sea a little more reliable and help explore and protect our marine environments have taken a leap forward through this advancement. With these enhanced images, we trained and tested a model designed to detect objects underwater. Our results demonstrated that this approach outperforms existing methods, achieving higher accuracy and producing more visually appealing results. This advancement not only makes underwater research more reliable but also aids in the exploration and protection of marine environments.**

***Index Terms*—Underwater Imaging, Salient Object Detection, Image Enhancement, USOD10K, Color Balance, Fusion**

I. Introduction

The underwater imaging itself is very important for marine biology, underwater archaeology, and autonomous underwater vehicles applications. The underwater images, however, are usually distorted by color cast, low contrast, and poor visibility due to light absorption and scattering while passing through the water medium. These issues make it difficult to achieve tasks such as underwater object detection and salient object detection.

To overcome these issues, the USOD10K dataset was proposed to establish a comprehensive benchmark for underwater salient object detection (USOD). The dataset screens from underwater images, covering 70 classes of saliency object in 12 types of underwater image scenes. For each image, salient object boundaries and depth maps are annotated, which offer a solid basis for training and evaluating USOD models.

At the same time, another method, introduced in ”Color Balance and Fusion for Underwater Image Enhancement”, provides an efficient method to enhance underwater images impaired by medium scattering and absorption. Furthermore, it does not need special devices, and prior information is not required about the underwater environment or nature of the scene. Its color balance and image fusion strategies work wonders to construct a more vivid image; thus, it acts as a valuable preprocessing task in the image dataset dedicated to underwater object detection models.

In this paper, we amalgamate this USOD10K dataset with a color balance and fusion enhancement process to enhance the performance of the typical underwater salient object detection. Specifically, we preprocess the USOD10K images with the color balance and fusion method to compensate for the frequent underwater image degradations and equipped our detection model.

This study aims to improve the performance of the salient object detection in underwater scenes by evaluating the advantages of coupling excellent image enhancement algorithms with large-scale manually labeled datasets. We hope to rectify the limitations found in the USOD10K dataset and also use color balance and fusion enhancement in order to together create even better data to make the most resilient possibility for underwater models.

II. Dataset

To facilitate research on underwater salient object detection (USOD), we introduce the USOD10K dataset, which serves as a large-scale and comprehensive benchmark. It consists of 10,255 underwater images, annotated with pixel-level salient objects representations, object boundaries, and predicted depth maps. The dataset covers 70 types of salient objects in 12 different underwater scenarios, which is a strong basis for training and evaluating the USOD models.

Construction Process:

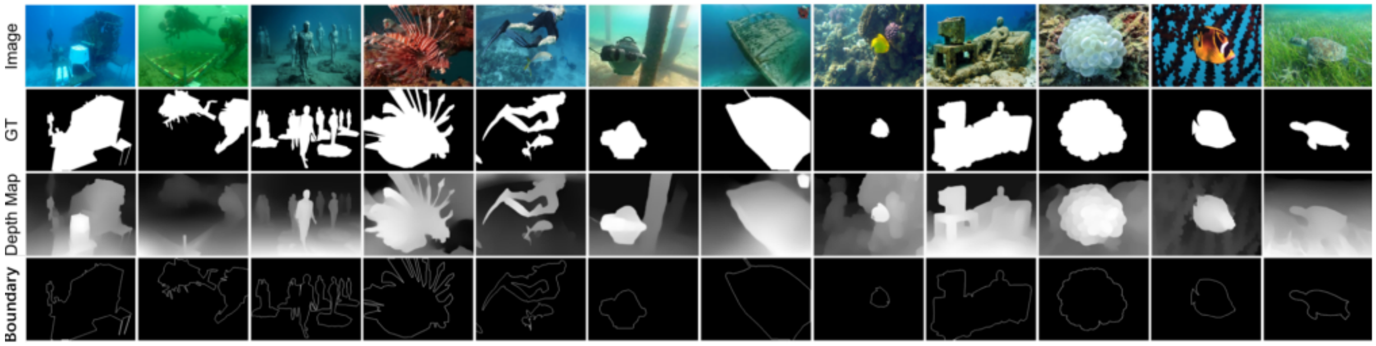

Fig. 1: Examples of the USOD10K dataset, showing input underwater images (Image), ground truth annotations (GT), estimated depth maps (Depth Map), and salient object boundaries (Boundary). This figure highlights the diversity and complexity of the dataset.

1) **Collection of Images**: More than 30,000 candidate underwater images were obtained from multiple sources, such as Google and Bing search engines, other datasets, and field images from studies of underwater engineering in the ocean, lake, and pool scenarios.
2) **Image Filtering**: A group of five trained volunteers manually filtered the images, removing duplicates and unusable files; 15,000 non-needle images remained for annotation.
3) **Image Annotation**: Eight professional annotators used pixel-wise tools to label the salient objects in each image. The annotation process was carried out ensuring consistency and correctness by using voting and cross-checking. The quality of annotations is further verified by two other USOD specialist volunteers. Out of this detailed procedure, we harvested 10,255 high-quality annotated images.
4) **Dataset Splitting**: In order to ensure efficient training and evaluation of the classifier, the dataset was split into training, validation, and test sets in the ratio 7:2:1, respectively. This separation results in uniform data distribution in various categories among the subsets.

Dataset Characteristics:

- **Salient Object Count**: 7,832 images contain a salient object; 1,701 contain 2 salient objects; 722 contain 3 or more salient objects. This distribution can be the foundation of research for both single salient object as well as multiple salient object detection problems.
- **Object Size**: The sizes of salient objects range from 0.05% to 93.98% with an average size of 14.12% (size defined as the ratio of object pixels to image pixels). Images in the dataset are categorized as large ($\geq 30\%$), medium (5%–30%), or small ($\leq 5\%$) with 1,357, 5,693, and 3,205 images in these categories, respectively.
- **Object Location**: The distribution of all samples indicates that salient objects have a bit of center bias and are more concentrated in the center area of the image.
- **Color Channel Intensity**: Consistent with underwater imaging characteristics, the red channel exhibits the lowest intensity due to greater absorption, while the green and blue channels are more prominent.
- **Object Categories**: The dataset's salient objects are organized into eight super-categories—obstacles, facilities, underwater animals, humans, relics, marine fish, plants, and litter—further divided into 70 sub-categories. This hierarchical structure reflects the diversity and complexity of underwater scenes.

Additional Features:

- **Depth Maps**: Given the challenges of obtaining depth information underwater, the dataset includes estimated depth maps for each image. Among several methods evaluated, the Dense Prediction Transformer (DPT) was selected for its superior performance in generating these depth maps.
- **Boundary Annotations**: Recognizing the importance of boundary information in salient object detection, the dataset provides boundary annotations for all salient objects, facilitating the development of models that leverage boundary cues.

The USOD10K dataset's extensive annotations and diverse content make it a valuable asset for advancing USOD research. It enables the development and evaluation of models capable of handling the unique challenges of underwater environments, such as varying object sizes, multiple salient objects, and complex backgrounds. By providing depth maps and boundary information, the dataset supports the creation of more sophisticated models that can integrate multiple sources of information to improve detection accuracy.

III. Preprocessing of Dataset

Preprocessing is a crucial initial step before utilizing the USOD10K dataset for training or testing underwater salient object detection models. Underwater images often suffer from issues such as color distortion, low contrast, and reduced visibility due to light absorption and scattering. To address these challenges, we employ a series of enhancement techniques, including color balance and fusion, structured as follows:

1. White Balancing: This step aims to correct color casts introduced by underwater lighting conditions, restoring natural

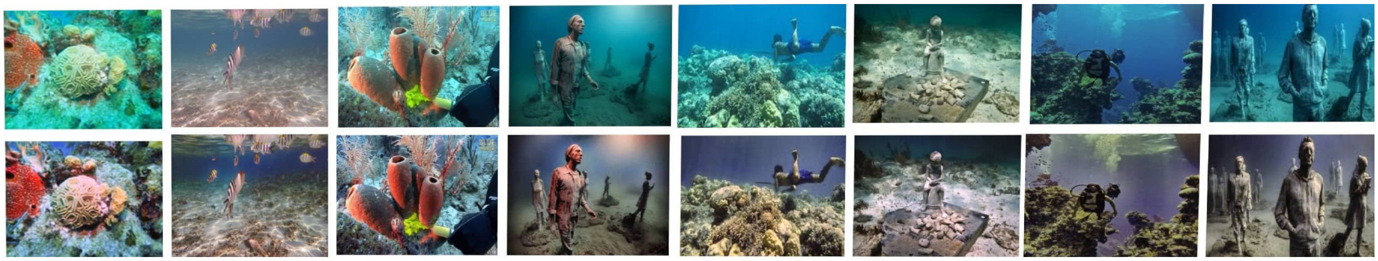

Fig. 2: Comparison of dataset inputs at various stages of preprocessing: The first row shows raw underwater images from the USOD10K dataset, highlighting issues like low contrast, color distortion, and visibility loss. The second row presents the results of the preprocessing step using the Color Balance and Fusion for Underwater Image Enhancement method, showcasing improved color balance, contrast, and visibility.

colors to the images. We apply the Gray World algorithm, which operates under the assumption that the average color of a scene under neutral lighting is gray. The algorithm adjusts each color channel (Red, Green, Blue) by scaling them based on their respective average intensities.

Gray World Algorithm:

Let R_{avg}, G_{avg}, and B_{avg} represent the average intensities of the Red, Green, and Blue channels, respectively. The scaling factors for each channel are computed as:

$$\text{Scale}_R = \frac{(R_{\text{avg}} + G_{\text{avg}} + B_{\text{avg}})/3}{R_{\text{avg}}}$$

$$\text{Scale}_G = \frac{(R_{\text{avg}} + G_{\text{avg}} + B_{\text{avg}})/3}{G_{\text{avg}}}$$

$$\text{Scale}_B = \frac{(R_{\text{avg}} + G_{\text{avg}} + B_{\text{avg}})/3}{B_{\text{avg}}}$$

Each pixel value in the Red channel is multiplied by Scale_R, in the Green channel by Scale_G, and in the Blue channel by Scale_B to achieve white balance.

2. Gamma Correction: Following white balancing, gamma correction is applied to adjust the luminance levels, enhancing details in darker regions and ensuring appropriate overall brightness. This nonlinear operation modifies the intensity values to improve visibility and contrast.

Gamma Correction Formula:

For an input intensity I_{in} and a chosen gamma value γ, the output intensity I_{out} is calculated as:

$$I_{\text{out}} = I_{\text{in}}^{\gamma}$$

A gamma value $\gamma < 1$ brightens the image, while $\gamma > 1$ darkens it.

3. Image Sharpening: To enhance fine details and edges, an edge-preserving filter is applied. This step is crucial for highlighting features important for accurate object detection.

Unsharp Masking:

The sharpening process involves subtracting a blurred version of the image from the original image and then adding the result back to the original image:

$$I_{\text{sharp}} = I_{\text{original}} + \alpha \times (I_{\text{original}} - I_{\text{blurred}})$$

Where:

- I_{sharp} is the sharpened image.
- I_{original} is the original image.
- I_{blurred} is the blurred version of the original image.
- α is a scaling factor that controls the strength of the sharpening effect.

4. Multiscale Fusion: The outputs from the gamma correction and sharpening steps are combined using a multiscale fusion strategy. This method integrates information at different scales, resulting in an image with improved contrast, color fidelity, and sharpness.

Laplacian Pyramid Decomposition:

Each processed image is decomposed into multiple frequency bands using Laplacian pyramids:

$$L_l = I_l - G(I_{l+1})$$

Where:

- L_l is the Laplacian at level l.
- I_l is the image at level l.
- G represents the Gaussian pyramid used for smoothing.

Fusion of Laplacian Pyramids:

The Laplacian pyramids of the processed images are fused using weight maps to combine the most significant features from each:

$$L_{\text{fused},l} = \sum_k W_{k,l} \times L_{k,l}$$

Where:

- $L_{\text{fused},l}$ is the fused Laplacian at level l.
- $W_{k,l}$ is the weight map for input k at level l.
- $L_{k,l}$ is the Laplacian of input k at level l.

Reconstruction of Fused Image:

The final enhanced image is reconstructed by summing the fused Laplacian pyramids across all levels:

$$I_{\text{fused}} = \sum_l L_{\text{fused},l}$$

By applying this preprocessing pipeline to the USOD10K dataset, we aim to mitigate common degradations found in underwater imagery. This enhancement ensures that the dataset

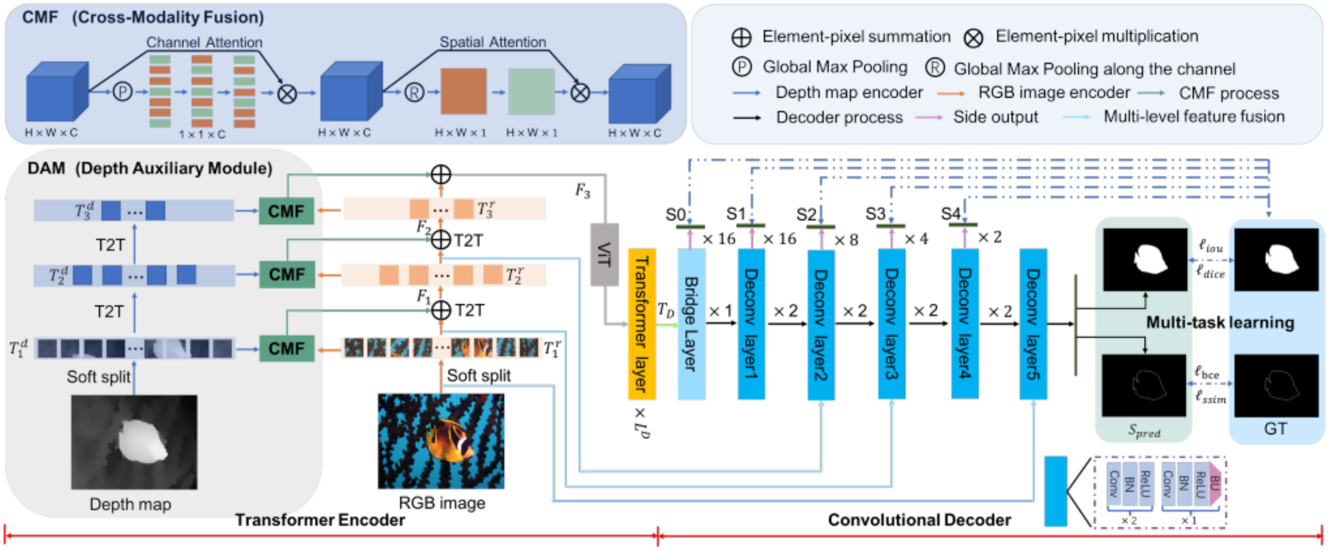

Fig. 3: Architecture of TC-USOD.

provides high-quality inputs for training and testing, thereby improving the performance and accuracy of the underwater salient object detection model.

The improved images are used to train the TC-USOD baseline model after preprocessing. The TC-USOD model introduced incorporates a hybrid architecture based on both encoder transformer and decoder convolutional networks, which allows it to learn global and local features that are comparably important in detecting salient objects in underwater scenes. By utilizing this hierarchical tokenization, the model learns multi-scale representations for detecting salient objects over different scales. The TC-USOD decoder includes several convolutional layers used to reverse the down-sampling process and create accurate high-resolution saliency maps. Skip connections form the decoder to bridge multi-level features from the encoder, enabling the model to effectively combine both low-level and high-level features for enhanced saliency map generation.

IV. Baseline Method

The new TC-USOD Model consists of four parts: transformer encoder, depth auxiliary model, convolutional decoder, and hybrid loss.

A. *Transformer Encoder*

Existing models in the field of object detection currently work on CNN where they focus on small localized regions and miss global dependencies. The new TC-USOD model addresses this by transformer-based architecture. The TC-USOD model uses T2T-ViT as the encoder to capture global dependencies. The Token-to-Token (T2T) module addresses the limitations of basic tokenization in ViT by iteratively tokenizing the image into smaller components and encoding local structural details using restructurization and soft-split mechanisms. During each iteration, the restructurization converts existing tokens into some new tokens and also integrates long-range dependencies in all tokens. The soft-split operation takes tokens in each $k \times k$ neighborhood and makes it into a new token for the next layer.

In this work, we follow to soft split input images into patch tokens and apply the T2T module twice on the patch tokens to obtain:

$$T_{r1} \in \mathbb{R}^{l_1 \times c}, \quad T_{r2} \in \mathbb{R}^{l_2 \times c}, \quad \text{and} \quad T_{r3} \in \mathbb{R}^{l_3 \times c}.$$

Token T_3 is further enhanced by using sinusoidal positional tokens to encode 2D position information. Then a L_d transformer is used to get long-range relationships in T_3, which results in another token $T_D \in \mathbb{R}^{l_3 \times d}$. After this, there is a connection between the transformer encoder and convolution decoder by concatenation. There is one more transformer encoder which processes depth maps to generate depth patch tokens T_{di} $(i = 1, 2, 3, 4)$ which are then fused with corresponding feature tokens T_{ri} $(i = 1, 2, 3, 4)$ to get F_i. The T2T final output uses ViT to adjust embedding dimensions from c to $d = 384$, after which L_d transformer extracts decoder patch tokens T_d.

B. *Depth Auxiliary Module*

In this work, a two-stream architecture is proposed, where depth maps generated by DPT are used to design a Depth Auxiliary Module (DAM). The DAM introduces the other stream to encode depth information, and then the corresponding feature maps are fused with those of the RGB stream to predict saliency masks. DAM employs the T2T module to extract multi-level depth patch tokens T_{di} from input depth maps, progressively fusing depth and RGB features. To remove the effect of low-quality depth maps, a Cross Modality Fusion (CMF) is integrated into the DAM, and this enhances the depth maps from both channel and spatial perspectives. These new depth features are added to the RGB features. Depth maps estimated for underwater images generally suffer from errors and noise, which in turn affects the working of the new TC-USOD model. To address this issue, a CMF strategy is designed to purify the estimated depth maps. Depth features T_{d1} are concatenated with their corresponding RGB features T_{r1}. The fused feature maps are then processed sequentially through channel attention and spatial attention mechanisms to purify the depth map. Finally, the refined depth features T_{di} are combined with T_{ri} to generate F_i ($i = 1, 2, 3$), which acts as input for the next T2T module. This process is mathematically expressed as:

$$F_i = T_{d_i} \otimes \mathrm{SA}(T_{d_i}) \otimes \mathrm{CA}(\mathrm{Cat}(T_{d_i}, T_{r_i})) \oplus T_{r_i}$$

where Cat represents concatenation followed by convolution, CA and SA are channel and spatial attention operations, $\otimes$ denotes element-wise multiplication, and $\oplus$ denotes element-wise addition.

C. *Convolutional Decoder*

Unlike other methods that use symmetrical encoder-decoder architecture, the new TC-USOD uses a convolutional decoder with a multi-level feature fusion strategy to decode patch tokens T_d into saliency maps. The decoder has six stages. The first stage acts as a bridge connecting the transformer encoder with the convolutional decoder. The next five stages share the same structure, where each stage comprises three convolutional layers, each followed by batch normalization and a ReLU activation function. The input for the 2nd, 3rd, and 5th deconvolution layers is formed by concatenating the upsampled output from the previous stage with the corresponding downsampled output from the encoder.

At each stage, the result undergoes a 3×3 convolutional layer, bilinear upsampling, and a sigmoid activation function to generate a side output saliency map. As there are six stages, we get six saliency maps during training, and at last, we get a final saliency map. Additionally, the new TC-USOD model uses a multi-task learning approach to simultaneously predict saliency and boundaries.

D. *Hybrid Loss*

A hybrid loss l is designed to guide the training process of the new TC-USOD model, which is given by the equation:

$$l = l_{\text{bce}} + l_{\text{iou}} + l_{\text{dice}} + l_{\text{ssim}}$$

where l_{bce}, l_{iou}, l_{dice}, l_{ssim} represent Binary Cross-Entropy loss, Intersection over Union loss, Dice loss, and Structural Similarity Index Measure loss, respectively. The BCE loss is the most widely used loss in binary segmentation. IoU is a standard evaluation measure for object detection and segmentation. Dice loss can evaluate the similarity between two samples based on the Dice coefficient. SSIM can assess image quality by capturing structural information in the image [**?**]. In this work, the SSIM loss is introduced to learn the boundary information of salient objects.

E. *Implementation Details*

The pretrained T2T-ViT$_{\text{t-14}}$ and ResNet34 are respectively adopted as the backbone of the transformer encoder and convolutional decoder in the new TC-USOD model. The new TC-USOD model uses the efficient Performer and $c = 64$ in T2T modules and sets $L_c = 14$. The batch size is set to 8, and the total training steps are 60,000. Adam is adopted as the optimizer. The learning rate is set to 0.0001. Multi-level feature fusion and multi-task learning strategy are adopted to generate binary maps and boundaries of salient objects at each convolutional decoder stage. The new TC-USOD model is implemented using PyTorch and trained on an NVIDIA GeForce GTX 1080 Ti GPU.

V. Experimental Settings

A. *Evaluation Metrics*

The evaluation metrics used in this model are Precision-Recall (PR) curve, S-measure (S_m), max E-measure ($E_{\max}$), max F-measure (maxF), and Mean Absolute Error (MAE).

VI. Ablation Studies

A. *Effectiveness of Hybrid Architecture*

The new TC-USOD model uses a new hybrid architecture which consists of transformer encoder and a convolutional decoder. To verify that this model is superior, we define the pure convolutional encoder-decoder architecture as Baseline1, the pure transformer encoder-decoder architecture as Baseline2, and the proposed hybrid architecture without the DAM module, multi-level feature fusion, and multi-task learning strategies as BaseTC-USOD, and our new model as new BaseTC-USOD. The table below shows the results.

TABLE I: Performance Comparison of TC-USOD with Various Baselines

Settings	S_m	$E_{\max}$	**maxF**	**MAE**
Baseline1	.8222	.9074	.7959	.0628
Baseline2	.8988	.9574	.8984	.0356
Base TC-USOD (Transformer-conv)	.9026	.9591	.8983	.0310
Base TC-USOD (DAM)	.8982	.9585	.8930	.0414
Base TC-USOD (DAM, CMF)	.9097	.9592	.8977	.0262
Base TC-USOD (DAM, CMF, F)	.9159	.9610	.9140	.0228
Base TC-USOD (DAM, CMF, F, S)	.9215	.9683	.9236	.0201
New Base TC-USOD	**.9116**	**.9561**	**.9087**	**.0238**

B. *Effectiveness of DAM*

The DAM module enhances the performance of the USOD model by integrating depth maps with RGB images through the Cross Modality Fusion (CMF) strategy. To evaluate its impact, we perform an ablation study where we first add depth maps to the Base TC-USOD model by simply fusing the depth feature map with the corresponding RGB feature map using an addition operation. As seen in Table I, the introduction of DAM and CMF has improved our results.

C. *Effectiveness of Multi-Level Feature Fusion and Multi-Level Supervision*

In the TC-USOD baseline, the feature maps F_i (for $i = 1, 2, 3$) are progressively integrated with the 2nd, 3rd, and 5th deconvolution layers to provide low-level, fine-grained feature maps. To assess the effectiveness of this approach, an ablation study is performed, and the results are shown in Table I. The results demonstrate that incorporating the multi-level feature fusion strategy within the convolutional decoder further enhances the TC-USOD model's performance. As seen in Table I (DAM, CMF, F, S), the Base TC-USOD benefits significantly from the multi-level supervision, leading to a notable performance improvement.

D. *Effectiveness of Hybrid Loss*

A hybrid loss is developed to help in generating accurate binary masks and boundaries for salient objects. Ablation studies are conducted to identify the effectiveness of each loss term. The results are shown below in Table II.

TABLE II: Performance Comparison of TC-USOD with Various Loss Functions

Settings	S_m	$E_{\max}$	**maxF**	**MAE**
l_{bce}	.9126	.9542	.9105	.0224
$l_{\text{bce}} + l_{\text{iou}}$	.9172	.9631	.9170	.0220
$l_{\text{bce}} + l_{\text{dice}}$	.9192	.9654	.9215	.0230
$l_{\text{bce}} + l_{\text{ssim}}$	.4420	.4956	.2573	.1403
$l_{\text{bce}} + l_{\text{iou}} + l_{\text{dice}}$	.9170	.9601	.9140	.0226
$l_{\text{bce}} + l_{\text{iou}} + l_{\text{ssim}}$	.9189	.9643	.9216	.0223
$l_{\text{bce}} + l_{\text{dice}} + l_{\text{ssim}}$	.9161	.9574	.9154	.0299
$l_{\text{bce}} + l_{\text{iou}} + l_{\text{dice}} + l_{\text{ssim}}$	**.9215**	**.9683**	**.9236**	**.0201**
New TC-USOD	.9087	.9561	.9147	.0238

We define the TC-USOD architectures related to TC-USOD, and the highlighted points in Table II are related to the new TC-USOD.

VII. Discussions

A. *Other Uses of USOD10K*

- **Underwater Image Enhancement**: There is a high need to have high-quality underwater images to explore deep seas. But it is not possible to get a high-quality image by only using a camera as there are many factors like refraction and dispersion. To overcome this, we need to use trained models to increase the quality of an image to a higher extent, and training a model requires datasets which are very scarce in today's world. As USOD10K contains 10,225 underwater images, we can use this dataset to train models, and USOD10K works as an ideal dataset for models working on underwater images.
- **Marine Creature Detection and Classification**: There are many creatures and a large amount of wildlife and plantations living underwater. These are under threat of extinction, and there is a great need to keep a record of them. A human being manually observing the videos of underwater cameras all the time is not possible, and we need computers to do this work. For that, we need to train our models on images which have underwater creatures residing in them. USOD10K contains 6,909 samples of such images and may be very helpful in training models in this area.
- **Diving People Detection**: Nowadays, robots have started going underwater along with humans, and detecting divers using vision methods is not possible because of the scattering of light. This can be overcome by deep-learning-based models which can be trained using the USOD10K dataset, which contains 801 samples of diving people.
- **Underwater Litter Detection**: There are many cases of throwing away litter into the seas, and this poses a great danger to the underwater environment. There is a great need to remove this underwater litter, and doing that manually consumes a lot of time and is also hazardous to the life of the person. So we can use autonomous underwater vehicles which can do this work, and they can be trained using the USOD10K dataset which contains 726 samples of underwater litter.
- **Underwater Co-Salient Object Detection**: Co-salient object detection aims to detect the co-occurring salient objects in multiple images. In the USOD10K, 2,423 samples can be used within co-salient object detection and 7,832 for salient object detection in a group of images.

B. *Applications of USOD*

- **Underwater Object Detection and Tracking**: With the growing use of various visual sensors in underwater environments, leveraging visual saliency for object detection and tracking has gained significant attention.
- **Underwater Navigation and Mapping**: Moving underwater is completely different when compared to moving in air, as there are many factors like absorption and scattering of light. Any automatic vehicle moving underwater can use USOD models to find salient objects and move accordingly without colliding with any objects or animals underwater.

VIII. Potential Future Developments

- **Depth Map Estimation**: Depth information serves as a valuable complement for resolving challenges like overlapping objects and varying viewpoints in SOD. Acquiring depth information in underwater environments using sensors commonly used on land is often not possible. Therefore, in most cases, depth map estimation of underwater images is an ill-posed problem.
- **Weakly/Self/Unsupervised Learning**: The success of deep SOD methods largely depends on the availability of large-scale, pixel-level annotated datasets, and most of these methods are trained in a fully supervised setting. But annotating data pixel-wise for every image is a very intensive process and requires a huge amount of time. To address this challenge, several weakly supervised, self-supervised, and unsupervised methods for SOD have been introduced.

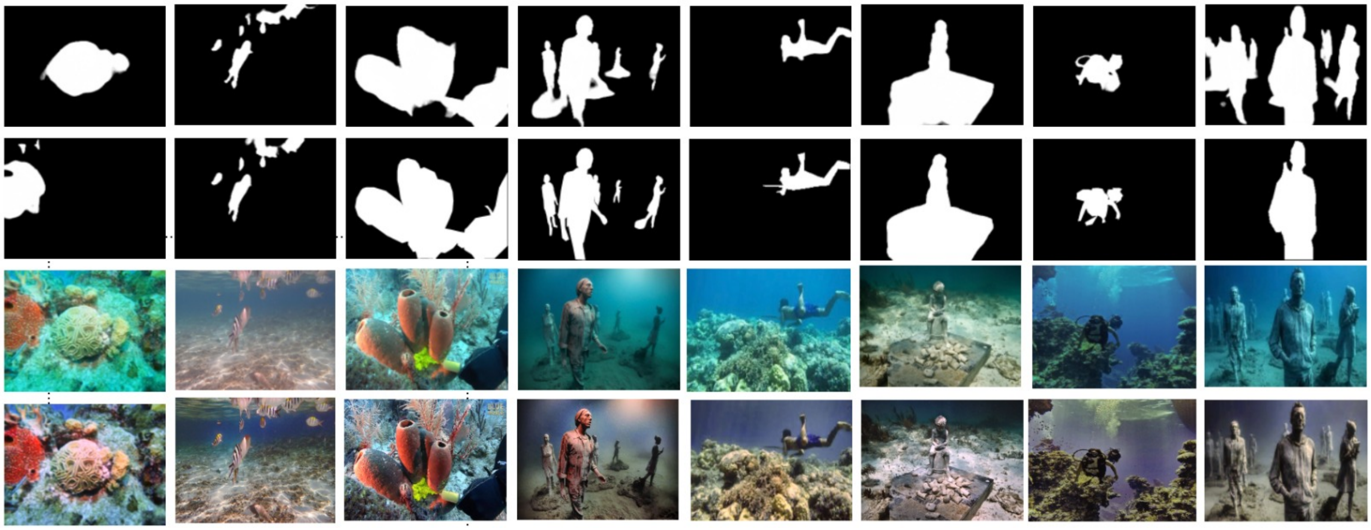

Fig. 4: Visualization of results across different processing stages: The first row shows the output saliency maps generated by the proposed New TC-USOD model, highlighting improved object boundaries and salient regions. The second row presents saliency maps generated by the baseline TC-USOD model, demonstrating its effectiveness but with less precision in object boundaries. The third row contains original images from the USOD10K dataset, showing the challenges of underwater imaging such as color distortion and scattering. The fourth row displays preprocessed images from the USOD10K dataset after applying the color balance and fusion enhancement, showing improved visibility, color fidelity, and contrast.

IX. Conclusions & Results

USOD10K is the first large dataset in the field of underwater exploration. This dataset contains 10,255 underwater images from 70 different categories and 12 different locations.

This dataset also contains depth maps and saliency of underwater objects. TC-USOD is a strong baseline in the field of underwater salient object detection and has given very good results.

To show how to make full use of the USOD10K dataset to design advanced USOD methods, a simple but strong baseline called TC-USOD is proposed. The TC-USOD is a novel hybrid architecture that consists of a transformer encoder and a convolutional decoder. It introduces the DAM module, multi-level feature fusion, and multi-task learning strategy to fully use the RGB underwater images, estimated depth maps, and salient object boundaries to generate accurate full-resolution saliency maps.

TABLE III: Comparison between TC-USOD and New TC-USOD

Settings	MAE	meanF	meanE	AP	AUC
TC-USOD	.0228	.9021	.9568	.8953	.9607
New TC-USOD	.0238	.8946	.9516	.8963	.9638

We have improved the dataset using color fusion models, and the new TC-USOD has shown some better results when compared to the TC-USOD model, which can be evidently seen from Table III, where we have got better E-measure, Average Precision (AP), and better Area Under Curve (AUC). Some of the parameters not included in previous tables are included in the table-3.

New TC-USOD has a lower E-measure when compared to TC-USOD, which indicates that New TC-USOD has better performance in capturing object boundaries when compared to TC-USOD.

New TC-USOD has higher Average Precision (AP) when compared to TC-USOD, which indicates that New TC-USOD is marginally better at correctly identifying relevant objects across various threshold values, which is a key factor in object detection tasks. Even though the difference is small, this shows New TC-USOD's superior ability to distinguish between salient and non-salient areas.

New TC-USOD has higher Area Under Curve (AUC) compared to TC-USOD, which indicates that New TC-USOD has a slightly better overall performance across different operating points. A higher AUC indicates that New TC-USOD is better at ranking positive instances higher than negative ones, improving its ability to distinguish between salient and non-salient regions.

Some of the final output images when tested through TC-USOD and New TC-USOD are shown in Figure 4.

We hope our work will boost the development of USOD research. By using advanced physics techniques we can further enchance under water images and we can remove all the effects of scattering and absorption. This helps in creating images which look like terrestrial images.This interdisciplinary collabaration has helped us in getting good result in a topic like USOD, which is a young research field and we beleive that through advanced techniques and interdisciplinary collabaration USOD can be done with a maximum accuracy.

X. Contributions

The main problem in USOD arises due to the nature of underwater images which are distracted due to physical phenomena like absorption and scattering of light. Also, some colors are not shown properly in underwater images.

We mainly worked to tackle this issue and we implemented a topic called color balance and fusion for underwater image enhancement, which removes the effect of blue light and many underwater scattering effects. Improved images after resolving issues can be seen in Fig-2.

The base TC-USOD model uses the USOD-10K dataset which directly takes underwater images and produces outputs.

We now enhanced the dataset using the above-mentioned techniques and we trained our model on this enhanced dataset. The model trained on the new dataset works better in USOD and this can be seen clearly in Fig-4.

The new TC-USOD detects boundaries of objects effectively when compared to the TC-USOD.

Further analyzing the obtained results from TC-USOD and new TC-USOD, we can see improvements in some of the error metrics, and this can be seen from Table-3 , Here we have got better E-measure, Average Precision (AP), and better Area Under Curve (AUC).

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 20–33, 2017.

[3] Y. Li, G. Liu, Q. Liu, Y. Sun, and S. Chen, "Moving object detection via segmentation and saliency constrained RPCA," *Neurocomputing*, vol. 323, pp. 352–362, 2019.

[4] R. Li, C.-H. Wu, S. Liu, J. Wang, G. Wang, G. Liu, and B. Zeng, "SDP-GAN: Saliency detail preservation generative adversarial networks for high perceptual quality style transfer," *IEEE Transactions on Image Processing*, vol. 30, pp. 374–385, 2021.

[5] Z. Ma, C. Wang, Y. Niu, X. Wang, and L. Shen, "A saliency-based reinforcement learning approach for a UAV to avoid flying obstacles," *Robotics and Autonomous Systems*, vol. 100, pp. 108–118, 2018.

[6] R. Mechrez, E. Shechtman, and L. Zelnik-Manor, "Saliency driven image manipulation," *Machine Vision and Applications*, vol. 30, no. 2, pp. 189–202, 2019.

[7] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[8] Y.-H. Wu, Y. Liu, J. Xu, J.-W. Bian, Y.-C. Gu, and M.-M. Cheng, "Mobilesal: Extremely efficient RGB-D salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[9] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: a benchmark and algorithms," in *European Conference on Computer Vision*. Springer, 2014, pp. 92–109.

[10] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2075–2089, 2021.

[11] M.-M. Cheng, S.-H. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[12] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

[13] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.

[14] M. D. Kocak, F. R. Dalgleish, M. F. Caimi, and Y. Y. Schechner, "A focus on recent developments and trends in underwater imaging," *Marine Technology Society Journal*, vol. 42, no. 1, pp. 52–67, 2008.

[15] G. L. Foresti, "Visual inspection of sea bottom structures by an autonomous underwater vehicle," *IEEE Transactions on Systems, Man, and Cybernetics B, Cybernetics*, vol. 31, no. 5, pp. 691–705, Oct. 2001.

[16] A. Ortiz, M. Simó, and G. Oliver, "A vision system for an underwater cable tracker," *Machine Vision and Applications*, vol. 13, pp. 129–140, Jul. 2002.

[17] A. Olmos and E. Trucco, "Detecting man-made objects in unconstrained subsea videos," in *Proceedings of the British Machine Vision Conference (BMVC)*, Sep. 2002, pp. 1–10.

[18] B. A. Levedahl and L. Silverberg, "Control of underwater vehicles in full unsteady flow," *IEEE Journal of Oceanic Engineering*, vol. 34, no. 4, pp. 656–668, Oct. 2009.

[19] C. H. Mazel, "In situ measurement of reflectance and fluorescence spectra to support hyperspectral remote sensing and marine biology research," in *Proceedings of IEEE OCEANS*, Sep. 2006, pp. 1–4.

[20] Y. Kahanov and J. G. Royal, "Analysis of hull remains of the Dor D Vessel, Tantura Lagoon, Israel," *International Journal of Nautical Archaeology*, vol. 30, pp. 257–265, Oct. 2001.

[21] R. Schettini and S. Corchs, "Underwater image processing: State of the art of restoration and image enhancement methods," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, Dec. 2010, Art. no. 746052.

[22] S. G. Narasimhan and S. K. Nayar, "Contrast restoration of weather-degraded images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 713–724, Jun. 2003.

[23] D.-M. He and G. G. L. Seet, "Divergent-beam LiDAR imaging in turbid water," *Optics and Lasers in Engineering*, vol. 41, pp. 217–231, Jan. 2004.

[24] Y. Y. Schechner and Y. Averbuch, "Regularized image recovery in scattering media," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1655–1660, Sep. 2007.

[25] R. Fattal, "Dehazing using color-lines," *ACM Transactions on Graphics*, vol. 34, Nov. 2014, Art. no. 13.

[26] H. Lu, Y. Li, S. Nakashima, H. Kim, and S. Serikawa, "Underwater image super-resolution by descattering and fusion," *IEEE Access*, vol. 5, pp. 670–679, 2017.

[27] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE CVPR*, Jun. 2015, pp. 5197–5206.

[28] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," in *Proceedings of the IEEE CVPR*, Jun. 2012, pp. 81–88.

[29] P. L. Drews, E. R. Nascimento, S. S. Botelho, and M. F. M. Campos, "Underwater depth estimation and image restoration based on single images," *IEEE Computer Graphics and Applications*, vol. 36, no. 2, pp. 24–35, 2016.

[30] L. Hong, X. Wang, Z. Xiao, G. Zhang, and J. Liu, "WSUIE: Weakly supervised underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8237–8244, 2021.

[31] H. Gupta and K. Mitra, "Unsupervised single image underwater depth estimation," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019, pp. 624–628.

[32] B. L. McGlamery, "A computer model for underwater camera systems," in *Ocean Optics VI*, S. Q. Duntley, Ed., vol. 0208, *International Society for Optics and Photonics*, SPIE, 1980, pp. 221–231.

[33] J. S. Jaffe, "Computer modeling and the design of optimal underwater imaging systems," *IEEE Journal of Oceanic Engineering*, vol. 15, no. 2, pp. 101–111, 1990.

[34] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1682–1691.

[35] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems*, vol. 27, 2014.