Enhancing Underwater Salient Object Detection with the Improved New TC-USOD Model

ALLURI LAKSHMAN NARENDRA-220010002

ATTUNURI Praneeth Reddy -220030005

Kapse Karthik- 220020026

PROBLEM STATMENT

Underwater environments pose challenges for object detection due to light absorption, color distortion, and blurry visuals, affecting marine research on coral reefs and marine animal tracking. Using the USOD10K dataset with preprocessing inspired by "Color Balance and Fusion for Underwater Image Enhancement," we enhanced underwater images by correcting color channels and reducing distortions. This enabled improved detection of marine objects, even in poor lighting and turbid conditions. Our trained model achieved higher accuracy and more visually appealing results compared to existing methods, advancing underwater research and aiding marine exploration and conservation.

DATASET CONSTRUCTION

- 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 crosschecking. 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 subset

DATASET CHARACTERSTICS

- **Salient Object Count**:
  - **7,832 images** with 1 salient object
  - **1,701 images** with 2 salient objects
  - **722 images** with 3+ salient objects
  - Supports research on both single and multiple salient object detection.

- **Object Size**:
  - Sizes range from **0.05% to 93.98%** (average: **14.12%** of image pixels).
  - Categories: **Large (≥30%)**: 1,357, **Medium (5%–30%)**: 5,693, **Small (≤5%)**: 3,205 images.

Object Location:

- Objects show **center bias**, concentrated near the image center.

- **Color Channel Intensity**:
  - **Red channel** is the weakest due to absorption, with **green** and **blue channels** being more prominent.

DATASET CHARACTERSTICS

Additional Features:-

- **Depth Maps**:
  - Estimated depth maps generated using **Dense Prediction Transformer (DPT)** for superior accuracy.
- **Boundary Annotations**:
  - Provides detailed boundary information for all salient objects, aiding boundary-focused model development.

Challenges Observed

Problems we observed: we observed that the outputs have more dominance of red and blue color’s blurred and unclear regions these may affect the detection accuracy

Impact on Outputs: Incorrect detection of salient objects due to color and clarity issues.

OUR APPROACH

Our approach addresses the challenges of underwater image degradation caused by light scattering and absorption through an innovative single-image enhancement pipeline. By leveraging a fusion-based strategy, we combine white-balancing techniques tailored to underwater conditions with multiscale fusion for artifact-free image blending. This process enhances edge sharpness, improves global contrast, and restores natural color tones without requiring specialized hardware or prior knowledge of scene structure. The robustness and effectiveness of our solution are validated through qualitative and quantitative assessments, proving its capability to significantly enhance underwater image quality and support downstream tasks such as segmentation and feature matching.

Tackling Colour Distortion and Contrast Issues

Preprocessing: White Balancing and Gamma Correction

- **White Balancing**:
  - Corrects underwater lighting colour casts to restore natural colours.
  - Utilizes the **Gray World Algorithm**:
    - Adjusts Red, Green, and Blue channels using scaling factors.
    - Applies scaling to pixel values for each channel.
- **Gamma Correction**:
  - Enhances visibility and details in darker regions.
    - $\gamma<1$: Brightens the image.
    - $\gamma>1$: Darkens the image.

Enhancing Sharpness and Fusion of Features

Preprocessing: Image Sharpening and Multiscale Fusion

- **Image Sharpening**:
  - Enhances fine details and edges in underwater images.
  - Preserves and highlights critical features for accurate analysis.
- **Multiscale Fusion**:
  - Combines details from sharpened and gamma-corrected images for optimal enhancement.
  - Integrates features at multiple levels to achieve sharper, high-quality output

Introduction to TC-USOD and Key Modules

Problem Statement:

•RGB images lack depth information, making it difficult to detect salient objects in challenging scenarios (e.g., underwater images).

•Depth maps provide structural and spatial clues to enhance saliency detection.

Introduction:

•TC-USOD is designed to integrate RGB and Depth inputs to improve salient object detection in underwater environments.

•Utilizes **Cross-Modality Fusion (CMF)** and **Depth Auxiliary Modules (DAM)** for modality interaction.

Proposed Solution:

•**Depth Auxiliary Module (DAM)**:

- Introduced to encode depth information effectively and integrate it with RGB features.
- Combines **Channel Attention (CA)** and **Spatial Attention (SA)** to focus on the most relevant features.

•**Cross-Modality Fusion (CMF)**:

•A combination of **channel attention** and **spatial attention** mechanisms to purify and enhance RGB features using depth inputs.

•CMF enhances the interaction between RGB and depth streams, ensuring complementary modalities are fused

Encoder Workflow with DAM and CMF

01.

Encoder Structure:
•RGB and depth backbones use **T2T-ViT-t-14** as the base architecture.
•Features are extracted hierarchically at multiple levels (low-level to high-level semantics).

02.

DAM Process:
1.Takes depth map features from the depth encoder and RGB features from the RGB encoder.
2.Applies:
1. **Channel Attention**: Highlights feature maps most relevant to saliency.
2. **Spatial Attention**: Focuses on spatially significant regions.
3.Outputs fused features, which are added back to RGB features to enhance them.

03.

Soft Splitting in the Encoder:
•Depth maps are softly split into hierarchical features at multiple levels:
- T1d,T2d,T3dT_1^d, T_2^d, T_3^dT1d,T2d,T3d represent features extracted at increasing levels of abstraction.

•These depth features are fused with RGB features using CMF at each encoder stage.

Key Mechanism:
•Multi-level depth-RGB fusion ensures that spatial details (low-level) and semantic details (high-level) are captured effectively.

Connection to Vision Transformer (ViT)

Outputs from Encoder:

•RGB encoder outputs F3(high-level RGB features).

•Depth encoder outputs F3d(high-level depth features).

Vision Transformer (ViT):

•Combines F3and F3d using multi-head self-attention to capture long-range dependencies and modality alignment.

•Transformer encodes global relationships between salient regions in RGB and depth maps.

Mechanism of Transformer

•**Input**: extracted Hierarchical feature embeddings from RGB and depth encoders(rgb_vit and dep_vit) through Vit backbone.

•**Process**:

- Multi-head self-attention attends to salient regions across RGB and depth modalities.
- Embeddings are processed through a series of transformer layers.

•**Output**:

- rgb_fea_16: Processed RGB features.
- depth_fea_16: Processed depth features.
- Spatial resolution: 16×16 times, embedding dimension: 384.

Decoder Workflow Semantic Integration and Up Sampling

- **Decoder Inputs**:
  - From **transformer** rgb_fea_16: Processed RGB features and depth_fea_16: Processed depth features.
  - Intermediate features from the **Encoder**: F1,F2,F3 and Fd1, Fd2,Fd3.
- **Semantic Workflow**:
  - **Bridge Layer**:
    - Reduces transformer output dimensions to match encoder feature sizes.
  - Hierarchical decoding through **De Conv layers**:
    - Deconv1 refines features using F3.
    - Deconv2 incorporates F2, progressively integrating lower-level spatial details.
    - Deconv3 and Deconv4 further refine using F1, ensuring high-resolution output.
- **Multi-Level Feature Fusion**:
  - Combines high-level semantics from the transformer with low-level details from the encoder at each stage.
  - Achieves a balance between semantic richness and spatial accuracy.

Key Process:

- Decoder upsamples the fused features step-by-step to generate high-resolution saliency maps.
- Each deconvolution layer refines features further, using skip connections from encoder outputs.

Output:

- Saliency map predictions and object boundary maps.

Decoder Workflow Semantic Integration and Up Sampling

Deconv Layers and Multi-Level Feature Fusion

Title: *Deconvolution and Multi-Level Feature Fusion in Decoder*

- **Deconv1 to Deconv5**:
  - **Deconv1**: Upsamples 16×16→32×32 using a 4×4 transposed convolution.
  - **Deconv2**: Refines 32×32→64×6432
  - **Deconv3**: Refines 64×64→128×128
  - **Deconv4**: Refines 128×128→256×256.
  - **Deconv5**: Outputs final saliency map (256×256×1) using 3x3convolution.
- **Activation Function**:
  - **ReLU**:
    - Applied after each convolution to introduce non-linearity.
    - Helps learn complex patterns and prevent gradient vanishing.
- **Feature Fusion**:
  - At each deconv stage, encoder features (F3,F2,F1F_3, F_2, F_1F3,F2,F1) are fused with upsampled features.
  - Enhances spatial resolution while retaining semantic information.

Hybrid Loss and Its Effectiveness

•**Hybrid Loss Definition**:
Total loss (ℓ)= ℓ bce+ ℓ iou+ ℓ dice+ ℓ ssim

01

•**Binary Cross-Entropy (BCE) Loss**: Pixel-wise accuracy for saliency prediction.

02

•**IoU Loss**: Encourages significant overlap between predictions and ground truth.

03

•**Dice Loss**: Balances precision and recall, effective for class imbalance.

04

•**SSIM Loss**: Focuses on structural similarity for sharp boundary prediction.

Improvements:
•Ablation studies confirm that combining all four losses yields superior performance on the **USOD10K dataset**.
•Hybrid loss ensures:
- Accurate binary saliency maps.
- Well-defined object boundaries.

ARCHITECHTURE

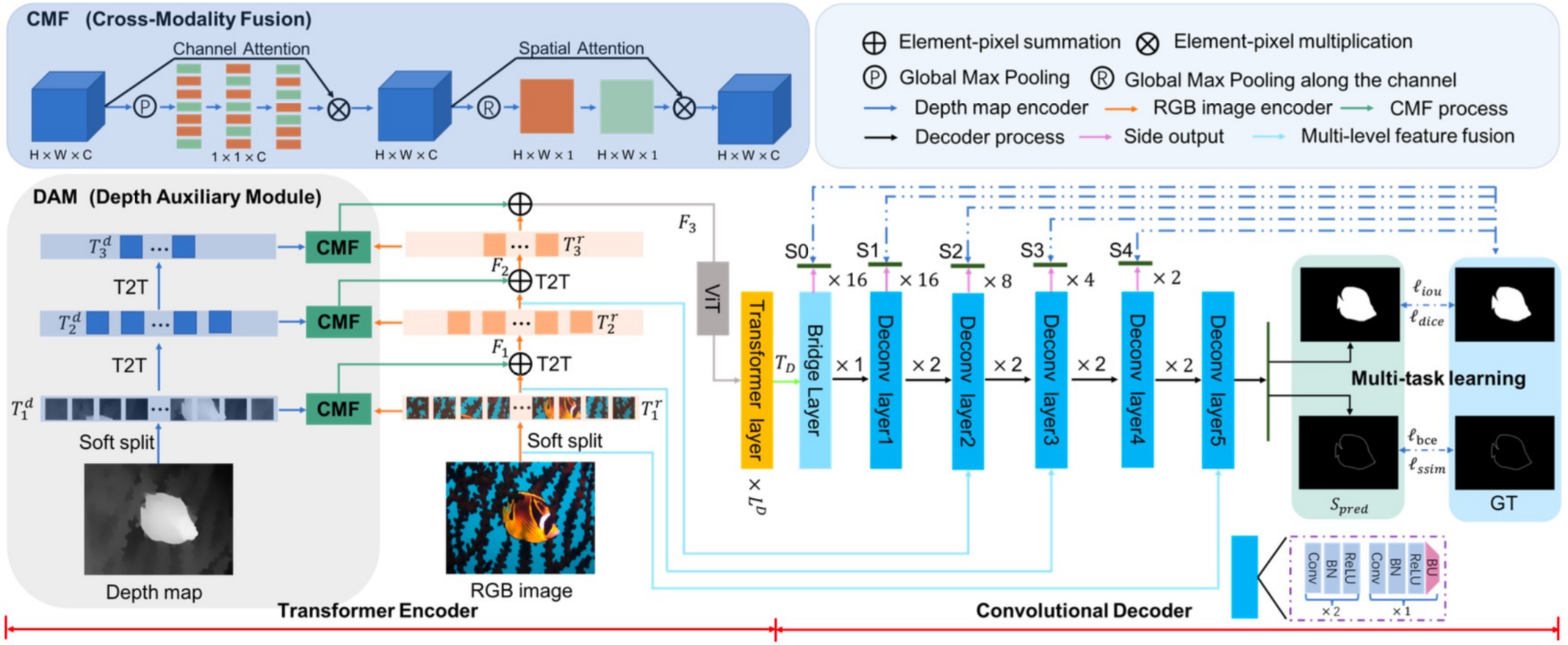


Fig. 6. **Overall architecture of the TC-USOD model**. The input RGB image and depth map are encoded by the T2T-ViT backbone [46], and then the saliency map and the salient object boundary are generated by a convolutional decoder with multi-level feature fusion and multi-task learning strategy. The DAM is designed to encode depth information, and the CMF module is to excavate helpful spatial clues from depth maps to purify RGB inputs.

Outputs Observed

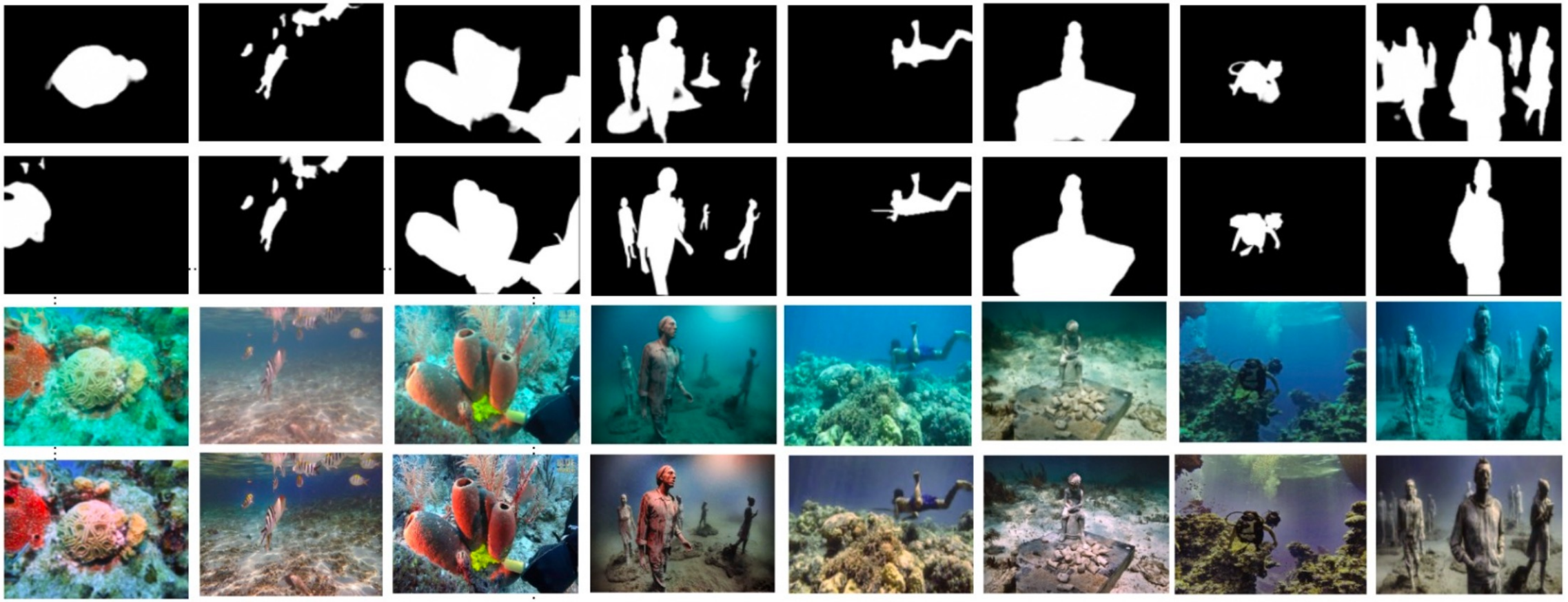


Fig. 4: Visualization of results across different processing stages: The first row shows the output saliency maps generated by the proposed New TC-USOD model, highlighting improved object boundaries and salient regions. The second row presents saliency maps generated by the baseline TC-USOD model, demonstrating 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.

Novelty and achievements

Changes made to improve the results

01. Major problem in USOD is nature of Underwater images which are significantly different when compared to terrestrial images.

02. So we applied some changes to underwater dataset and made it look like some terrestrial images with help of colour fusion and balance techniques.

03. This technique helps in detecting some of the small objects and colours which go unnoticed when we take underwater images.

04. With this techniques we have achieved marginal improvements in AUC , AP and Mean Error.

RESULTS

Settings	MEANE	MEANF	AP	MAE	AUC
TC-USOD	0.9568	0.9021	0.8953	0.0228	0.9607
New TC-USOD	0.9516	0.8946	0.8963	0.0238	0.9638

New TC-USOD performed better in terms of Area Under Curve , Mean Error and Average precision And the improvements in terms of results can be Seen in upcoming slides

Discussion On Results

E Measure

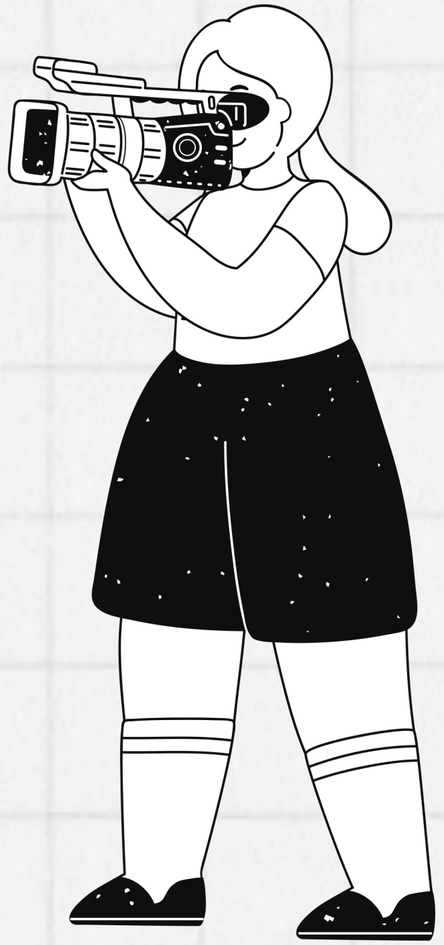
Lower Emeasure indicates that New TC-USOD has better performance in capturing object boundaries when compared to TC-USOD.

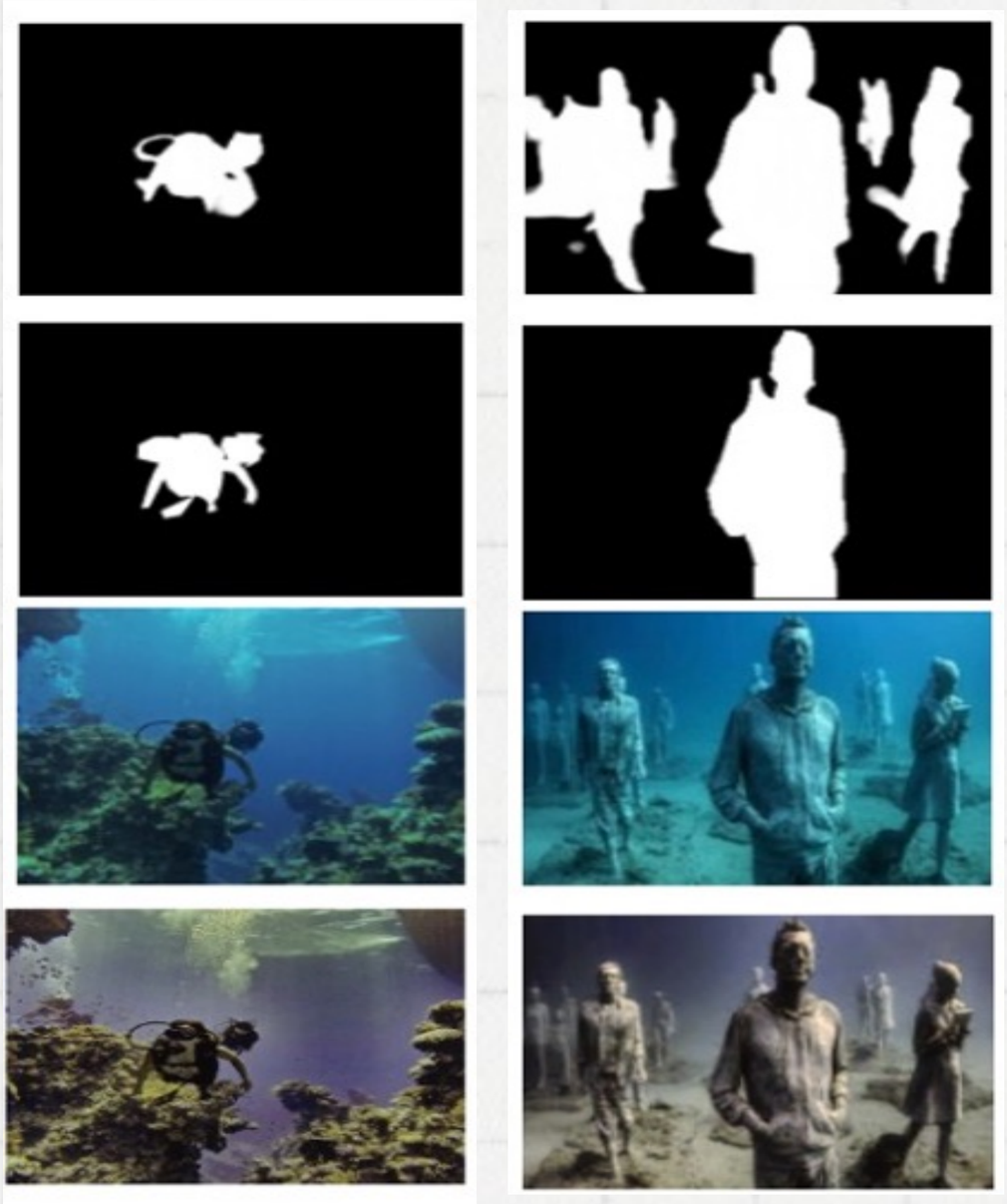
Average Precision

Higher Average precision indicates that New TC-USOD is better at correctly identifying relevant objects across various threshold values.

Area Under Curve

Higher Area Under Curve indicates that new TC-USOD is better at ranking positive than negatives when compared to TC-USOD. So, new TC-USOD has potential to work at different operating points.





Output from New TC-USOD Model

Output from TC-USOD Model

Image from USOD-10K dataset

Image from the pre-processed dataset

Future Scope

Exploring USOD

Depth Map Estimation

depth information can be re graded as complementary guidance to address the overlapping and viewpoint issues in SOD. it is impractical to obtain depth information of the observed underwater scene by using sensors that commonly used in the terrestrial field.

Mutual transfer between SOD and USOD methods

It is labor-costs and time-consuming to retrain models related to SOD and USOD. Thus we need to consider techniques through which we can transform the knowledge in between different SOD and USOD models.

Weakly/Self/Un-Supervised Learning

SOD methods relies on the availability of large-scale pixel-wise annotated datasets.making pixel wise annotations for training datasets is labor-intensive and time-consuming. This problem can be solved by using Weakly/Un-Supervised Learning.

Real-time USOD Inference

existing deep learning-based USOD methods cannot perform real time interference since they relay on high amount of computations.So it is difficult in case of robots moving under water to be autonomous.We can work and try to improve this.

CONCLUSIONS

- TC-USOD, a hybrid model using a transformer encoder and convolutional decoder, introduces the DAM module, multi-task learning, and feature fusion to generate accurate saliency maps.
- The enhanced New TC-USOD model shows improvements in Average Precision (AP) and Area Under Curve (AUC), excelling in distinguishing salient regions despite a marginally lower E-measure.
- Advanced physics techniques can further enhance underwater images by reducing scattering and absorption, fostering interdisciplinary research and propelling underwater salient object detection (USOD) advancements.

CONTRIBUTIONS

1. **Lakshman**:
    1. Identified the core challenges in underwater salient object detection (USOD), such as **light absorption**, **scattering effects**, and **colour distortion** in underwater images.
    2. Researched and implemented the **colour balance and fusion techniques** for underwater image enhancement.
    3. Focused on improving image quality by removing the **effects of blue light** and mitigating underwater scattering effects.
2. **Praneeth**:
    1. Enhanced the USOD-10K dataset by applying **Lakshman's colour balance and fusion techniques** to underwater images.
    2. Trained the **base TC-USOD model** on the enhanced dataset and validated its performance.
    3. Conducted comparative analysis between the base TC-USOD and the enhanced TC-USOD models.
    4. Highlighted improvements in salient object boundary detection with the enhanced model
3. **Karthik**:
    1. Performed detailed evaluation of the **error metrics**, comparing the performance of both models.
    2. Analyzed improvements in key metrics such as **E-measure**, **Average Precision (AP)**, and **Area Under the Curve (AUC)** (referenced in **Table above**).

Thank you