2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10611543

CaveSeg: Deep Semantic Segmentation and Scene Parsing for Autonomous Underwater Cave Exploration

Adnan Abdullah[1], Titon Barua[2], Reagan Tibbetts[2], Zijie Chen[1], Md Jahidul Islam[1], and Ioannis Rekleitis[2]

***Abstract*— In this paper, we present CaveSeg - the first visual learning pipeline for semantic segmentation and scene parsing for AUV navigation inside underwater caves. We address the problem of scarce annotated training data by preparing a comprehensive dataset for semantic segmentation of underwater cave scenes. It contains pixel annotations for important navigation markers (*e.g.* caveline, arrows), obstacles (*e.g.* ground plain and overhead layers), scuba divers, and open areas for servoing. Through comprehensive benchmark analyses on cave systems in USA, Mexico, and Spain locations, we demonstrate that robust deep visual models can be developed based on CaveSeg for fast semantic scene parsing of underwater cave environments. In particular, we formulate a novel transformer-based model that is computationally light and offers near real-time execution in addition to achieving state-of-the-art performance. Finally, we explore the design choices and implications of semantic segmentation for visual servoing by AUVs inside underwater caves. The proposed model and benchmark dataset open up promising opportunities for future research in autonomous underwater cave exploration and mapping.**

I. Introduction & Background

Underwater cave formations, sediment properties, and water chemistry provide insights into the world's past climate conditions and geological processes [1], [2]. Underwater caves also play a crucial role in monitoring and tracking groundwater flows in Karst topographies, while almost 25% of the world's population relies on Karst freshwater resources [3]. Despite the importance, underwater cave exploration and mapping by humans is a tedious, labor-intensive, and extremely dangerous operation, even for highly skilled divers [4]. When a new section of a cave is discovered, a single and continuous line termed *caveline* [5] is laid out identifying the skeleton of the main passages. The caveline is attached to other navigation guides such as *arrows* and *cookies* [6], marking the orientation of the cave, distance to the entrance and presence of other divers. Such surveys by the explorers produce a one-dimensional retraction of the 3D environment. Recording all this information together with additional observations [7] is a challenging, time-consuming, and error-prone process.

Enabling Autonomous Underwater Vehicles (AUVs) and Remotely Operated Vehicles (ROVs) to navigate, explore, and map underwater caves safely and effectively is of significant importance [2], [8]; Fig. 1(a) shows an ROV deployment scenario inside the Orange Grove Cave System in Florida.

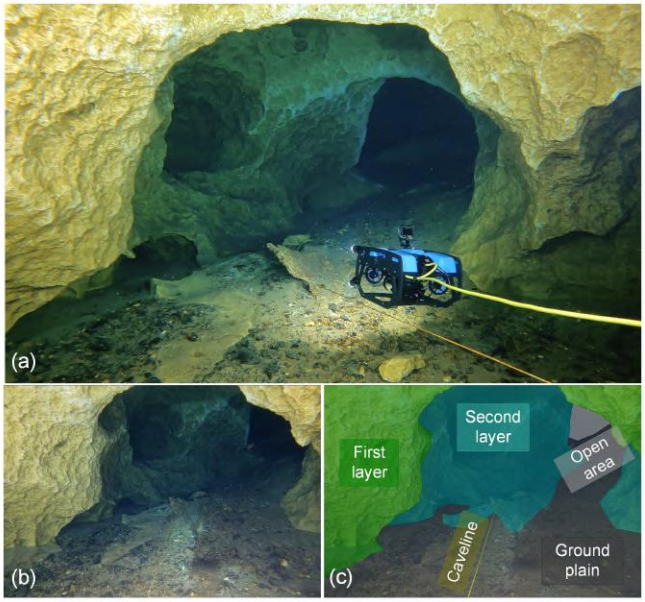

Fig. 1: (a) A tethered BlueROV2 is operating inside an underwater cave system in FL, USA; it is teleoperated by a surface operator following the *caveline* as a navigation guide; (b) the corresponding POV from the robot's camera; (c) the proposed semantic parsing concept is shown; the envisioned capabilities are: first-layer & second-layer obstacle avoidance, ground plain estimation, and caveline detection, following, and 3D estimation – to enable autonomous robot navigation inside underwater caves.

Our earlier work [9] developed a high-precision camera trajectory estimation method by a Visual-Inertial Odometry (VIO) algorithm [10], which can generate 3D caveline trajectory estimates comparable to the manually surveyed measurements. More recently in [6], we addressed the lack of annotated samples in visual learning for caveline detection and tracking across different scenes for AUV navigation.

In this work, we focus on developing a deep visual learning pipeline to extract dense semantic information for autonomous underwater cave exploration by mobile robots. Considering the limited onboard resources available on embedded platforms, our objective is to design a computationally light model that can (learn to) identify the navigation markers of underwater caves (*e.g.* caveline, arrows), obstacles to avoid (*e.g.* ground plain and overhead layers), scuba divers (for cooperative missions), and safe open areas for visual servoing in real-time. We identify two major difficulties to achieve these: (**i**) no large-scale datasets are available for underwater cave environments; (**ii**) the state-of-the-art (SOTA) models for semantic scene parsing are computationally too demanding for robotic platforms.

We address these challenges by proposing **CaveSeg**, the

[1]RoboPI Laboratory, Dept. of ECE, University of Florida, US. Email: {adnanabdullah@,chen.zijie@,jahid@ece.}ufl.edu.
[2]AFRL, Dept. of CSE, University of South Carolina, USA. {baruat@email,rbt@email,yiannisr@cse}.sc.edu.

first large-scale semantic segmentation dataset and learning pipeline for underwater cave exploration. We collect comprehensive training data by ROVs and scuba divers through robotics trials in three major locations [6]: the Devil's system in Florida, USA; Dos Ojos Cenote in QR, Mexico; and Cueva del Agua in Murcia, Spain. Our processed data contain 3350 pixel-annotated samples with 13 object categories that include first and second layer obstacles, human scuba divers, and navigation aids (see Sec III; Fig. 2). We also compile a **CaveSeg-Challenge** test set that contains 350 samples from unseen waterbody and cave systems such as the Blue Grotto and Orange Grove cave systems in FL, USA. We conduct extensive benchmark evaluation of SOTA models across the Convolutional Neural Network (CNN) [11], [12], Conditional Random Fields (CRF) [13], [14], and Vision Transformer (ViT) [15]–[17] literature, which validate that robust semantic learning is feasible on CaveSeg.

Moreover, we develop a novel **CaveSeg model** by rigorous design choices to balance the robustness-efficiency trade-off [18]. The proposed model consists of a transformer-based backbone, a multi-scale pyramid pooling head, and a hierarchical feature aggregation module for robust semantic learning (see Sec.IV). Experimental evaluation and comparison with SOTA models suggest that the CaveSeg model is over $3\times$ more memory efficiency and offers $1.8\times$ faster inference than SOTA models, while providing comparable benchmark performances. A series of experiments on unseen challenging scenes prove its robustness across different waterbody conditions and optical artifacts (see Sec.V).

Furthermore, we highlight several challenging scenarios and practical use cases of CaveSeg for real-time AUV navigation inside underwater caves. Those scenarios include safe cave exploration by caveline following and obstacle avoidance, planning towards caveline rediscovery, finding safe open passages and exit directions inside caves, giving uninterrupted right-of-way to scuba divers exiting the cave, and 3D semantic mapping and state estimation. We demonstrate that CaveSeg-generated semantic labels can be utilized effectively for vision-based cave exploration and semantic mapping by AUVs (see more in Sec. VI).

II. Related Work

A. Underwater Cave Exploration and Mapping

Exploration and mapping of underwater caves by human divers traditionally employed photogrammetry [19] methods in order to generate informative photorealistic representation, especially for sites with archaeological interest [1], [20]. Autonomous underwater cave mapping by AUVs remains an open problem due to many challenges. Mallios *et al*. [21] manually moved an AUV collecting acoustic data for offline mapping, and Weidner *et al*. [22], [23] used images from a stereo camera to create a 3D reconstruction of the cave walls, floor, and ceiling. Major challenges to vision-based state estimation in an underwater environment include lighting variations, light absorption, and blurriness [24]. Rahman *et al*. [10], [25] proposed a framework where visual, acoustic, inertial, and water depth data are fused together to estimate the trajectory of an AUV or a sensor while in parallel generates a sparse representation of the underwater cave. Denser models of the cave boundaries can be obtained by mapping the moving shadows [26], the contours [27], or via dense online stereo reconstruction [28].

More recently, Richmond *et al*. [2] describe a man-portable AUV named *Sunfish*, which can work safely inside underwater caves and bring back chemical profiles, detailed imagery, and sonar maps of the cave. Our recent work [6], [18] develops a ViT-based caveline detection model for autonomous caveline following by underwater robots. However, beyond detecting cavelines, a full-form scene parsing is essential for safe AUV navigation inside underwater caves.

B. Semantic Segmentation of Underwater Scenes

Deep visual learning algorithms have revolutionized semantic segmentation and scene parsing benchmarks on standard datasets, which are mostly developed for terrestrial applications. The existing learning pipelines trained on terrestrial imagery are not directly applicable because the object categories and image statistics are entirely different. A unique set of underwater image distortion artifacts and the unavailability of large-scale annotated datasets have influenced a significant lack of research attempts on semantic segmentation of underwater imagery. The contemporary research literature offers application-specific image datasets for coral reef segmentation and coverage estimation [29], visual attention modeling [30], human-robot cooperative missions [31], image restoration and foreground enhancement [32], and fish detection [33], [34].

More recent work by Modasshir *et al*. combined a deep learning-based classifier model to identify and track the locations of different types of corals to generate semantic maps [35] as well as volumetric models [36]. Islam *et al*. formulated the SUIM dataset [31] for semantic segmentation of underwater imagery with eight object categories: fish, coral reefs, aquatic plants, wrecks/ruins, human divers, robots/instruments, and sea-floor. Other datasets consider even fewer object categories such as marine debris or ship hull defects [37]. With limited training samples per object category over only a few waterbody types, it is extremely challenging to achieve good generalization performance by SOTA deep learning-based models for image segmentation and scene parsing. More importantly, these object categories are not useful for underwater cave exploration and mapping applications – which we address in this paper.

III. CaveSeg Dataset: Data Preparation and Problem Formulation

We prepared learning pipelines with data collected from **three cave systems** in different geographical locations [6]: the Devil's system in Florida, USA; Dos Ojos Cenote in QR, Mexico; and Cueva del Agua in Murcia, Spain. For the semantic labels, we considered the following object categories: caveline, first layer (immediate avoidance areas), second layer (areas to be avoided subsequently), open area

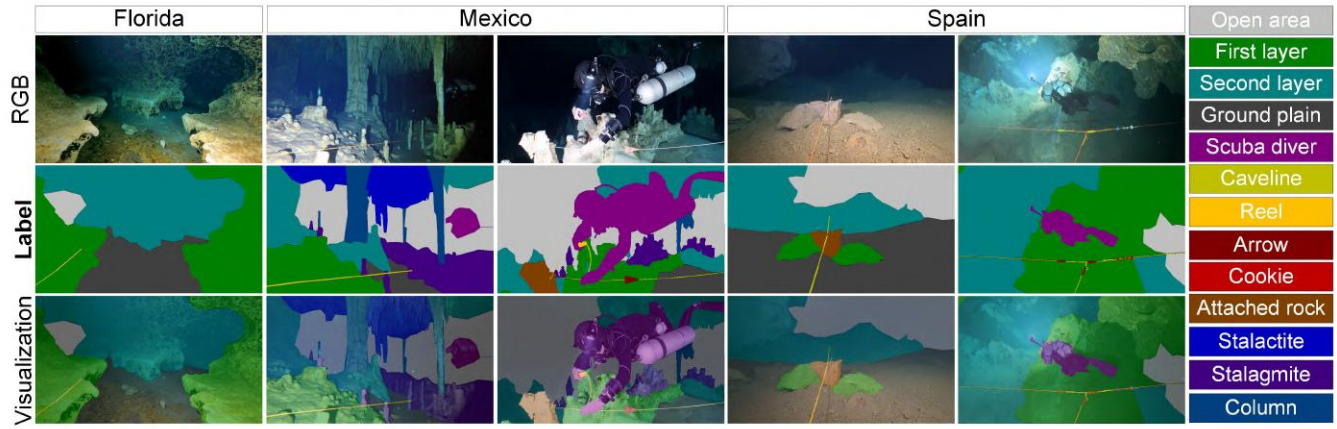

Fig. 2: A few sample images from the proposed **CaveSeg dataset**, corresponding ground truth labels, and their overlayed visualizations are shown; color codes for each object category are listed on the right.

(obstacle-free regions), ground plain, scuba divers, navigation aids (*e.g.*, arrows, reels, and cookies), and caveline-attached rocks. Moreover, cave ornaments (*e.g.*, stalactites, stalagmites, and columns) are also considered if they are present in the scene, which is the case for the Mexico caves. With these 13 object categories in consideration, a total of 3350 images are labeled in the **CaveSeg dataset**. A few annotated samples are shown in Fig. 2; the dataset and relevant information are available online at `https://robopi.ece.ufl.edu/caveseg.html`.

As mentioned earlier, the role of the caveline is crucial for any underwater cave operations. It represents the direction of exploration, the main area where the cave extends, and equally important, the path to safely exit the cave. Caveline pixels are marked as yellow in order to generate maximum contrast. The rock formations where the caveline is attached are often called *placements* or *attachment points*. These rocks most of the time signal a change in the direction of the caveline and they are marked in brown. Navigational markers such as *arrows* (dark red) and *cookies* (red) provide important information about the direction to the nearest cave exit and the presence of other divers in the cave, respectively.

A special category in our semantic mapping scheme is the scuba divers (magenta). An AUV should always give the right of way to divers, especially those exiting the cave. In case of an emergency, there should be nothing impeding the divers from reaching the surface, which is a norm practiced by cave divers. As such we have established a human diver category to incorporate emergency responses when a diver is detected, for example, lowering the light intensity, moving away from the main passage, avoiding abrupt motions, etc.

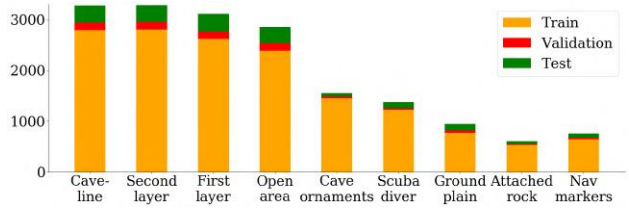

Fig. 3: Frequencies and distributions of important object categories are shown in the *train*, *validation*, and *test* sets.

Fig. 3 shows the frequency and distribution of various object categories in CaveSeg dataset. Human diver is present in 40% of the images. Over 90% of the samples contain caveline, obstacle-free open areas, as well as the first and second layer obstacles. Navigation markers (*e.g.*, cookies, arrows, reels) typically occupy smaller pixel areas compared to other objects, and they are found in 20% of the data. Overall, these 13 object categories in the scene embed useful information for vision-based planning and navigation for AUVs inside underwater caves.

IV. CaveSeg Model: Semantic Scene Segmentation Of Underwater Caves

A. Network Design and Learning Pipeline

We search for a computationally light architecture that provides real-time underwater cave scene segmentation in addition to achieving state-of-the-art (SOTA) performance. To this end, we explore both CNN-based and transformer-based backbone architectures [15], [38], [39]. The windowed multi-head self-attention (W-MSA) module proposed in Swin Transformer [17] is a powerful tool to maintain efficient computation. Additionally, the window shifting technique connects features across spatially overlapped windows. We find this connection to be particularly useful since cave scenes hold some spatial relation among categories. For instance, navigation markers and attachment rocks are almost always found on the caveline, while the second layer obstacles are usually found around the first layer or ground plain. Although computationally heavy, we found such cross-category attention extraction to be effective in cave scene segmentation tasks. Inspired by this, we design a novel architecture that incorporates these capabilities with a light Swin Transformer backbone for feature extraction. With an efficient Pyramid Pooling Module (PPM) and hierarchical feature aggregation, our proposed CaveSeg model is over $3.3\times$ memory efficient and offers 60% faster inference rates than the Swin Transformer base model.

1) CaveSeg Model Architecture: The detailed network architecture is shown in Fig 4. First, input RGB images are partitioned into 4×4 non-overlapping patches or `tokens`; a linear embedding layer of dimension 48 is then applied

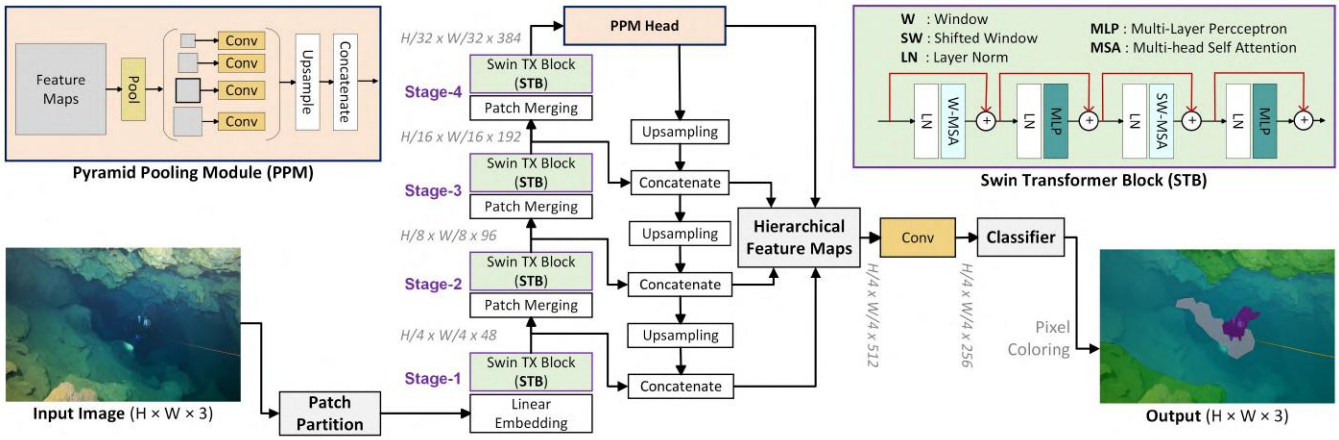

Fig. 4: The network architecture of our proposed **CaveSeg model** is shown. Input images are partitioned into 4×4 patches and fed into a four-stage transformer backbone for coarse-to-fine feature extraction. The extracted multi-scale features are then pooled and combined by the PPM head for bottom-up and top-down feature aggregation. A hierarchical feature map is then compiled by merging several multi-level feature representations. On this feature space, a classifier performs pixel-level semantic segmentation to generate the final outputs.

to each token. These feature tokens are passed through a four-stage backbone, each containing a windowed and shifted windowed module of multi-head self-attention [40]. In each stage, patches are merged with 2×2 neighboring regions to reduce the number of tokens while at the same time, the linear embedded dimension is doubled. Subsequently, bottom-up and top-down feature aggregation [41] is performed in two separate branches. A pyramid pooling module (PPM) [42] is attached to the backbone that further improves global feature extraction at deep layers of the network [43]. Features from each stage of the backbone as well as from the PPM head are then fused and compiled into a hierarchical feature map. Finally, a fully connected convolution layer performs pixel-wise category estimation on that feature space.

2) Supervised Learning Pipeline: The end-to-end training is driven by the standard cross-entropy loss [44] that quantifies the dissimilarity in the generated and predicted pixel labels for each category. For multi-class segmentation of an image with N pixels and M classes, it is calculated as

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_{i,c}\log(p_{i,c}), \tag{1}$$

where $p_{i,c}$ denotes the probability of pixel i belonging to class c, and y is 0 (or 1) for correct (or incorrect) predictions, respectively. The training is optimized by the Stochastic Gradient Descent (SGD) algorithm with an initial learning rate of $1e^{-4}$ and a momentum of 0.9.

B. Training Setups for CaveSeg and Baseline Models

We configured a unified training pipeline for CaveSeg and several SOTA benchmark models for training and evaluation. Specifically, we used the MMSegmentation libraries [45] in PyTorch for the large-scale training on CaveSeg dataset with a 85:5:10 split ratio for train, validation, and test, respectively. For baseline comparison, we considered the following SOTA models across the CNN, CRF, and ViT literature: FastFCN [39], DeepLabV3+ [38], Segmenter [16], Segformer [15], and Swin Transformer [17]. The proposed CaveSeg model is pre-trained on ADE20k [46] followed by the unified training. The input-output resolution is set to 960×540 for all models; other SOTA model-specific parameters are chosen based on their recommended configurations.

TABLE I: Semantic segmentation performances of all models are compared on 350 test images from the CaveSeg-Challenge set; here, higher scores (↑) are better for all metrics in consideration.

Method	mIoU ↑	mAcc ↑	aAcc ↑
FastFCN [39]	38.86	46.89	72.01
DeepLabV3+ [38]	38.46	49.47	71.64
Segmenter [16]	30.81	39.64	69.76
Segformer [15]	35.36	44.71	70.19
Swin Transformer [17]	**48.11**	**56.69**	**73.26**
CaveSeg (proposed)	40.22	45.99	72.91

V. Performance Analyses of CaveSeg

A. Quantitative Evaluation

We use three standard metrics for quantitative assessments: mean Intersection Over Union **mIOU**, mean class-wise Accuracy (**mAcc**), and Average pixel Accuracy (**aAcc**). The IoU (Intersection Over Union) measures caveline localization performance using the area of overlapping regions of the predicted and ground truth labels, while mAcc and aAcc represent the mean accuracy of each class category and over all pixels, respectively. The quantitative results are in Table I; all evaluations are performed on the **CaveSeg-Challenge** set, which we curated with 350 test samples. These samples include low-light noisy scenarios in the CL-Challenge set [6], as well as data from cave explorations in Blue Grotto and Orange Grove cave systems in FL, USA.

As Table I shows, our proposed dataset and learning pipelines can achieve up to 73% accuracy for pixel-wise segmentation. The proposed CaveSeg model offers comparable performance margins despite having a significantly lighter architecture. The comparisons for computational efficiency are provided in Table II. CaveSeg model has less than 50% parameters than other competitive models and offers up to $1.8\times$ faster inference rates. While we analyze the class-wise performance, we find that small and rare object categories such as arrows, cookies, and reels are challenging in general.

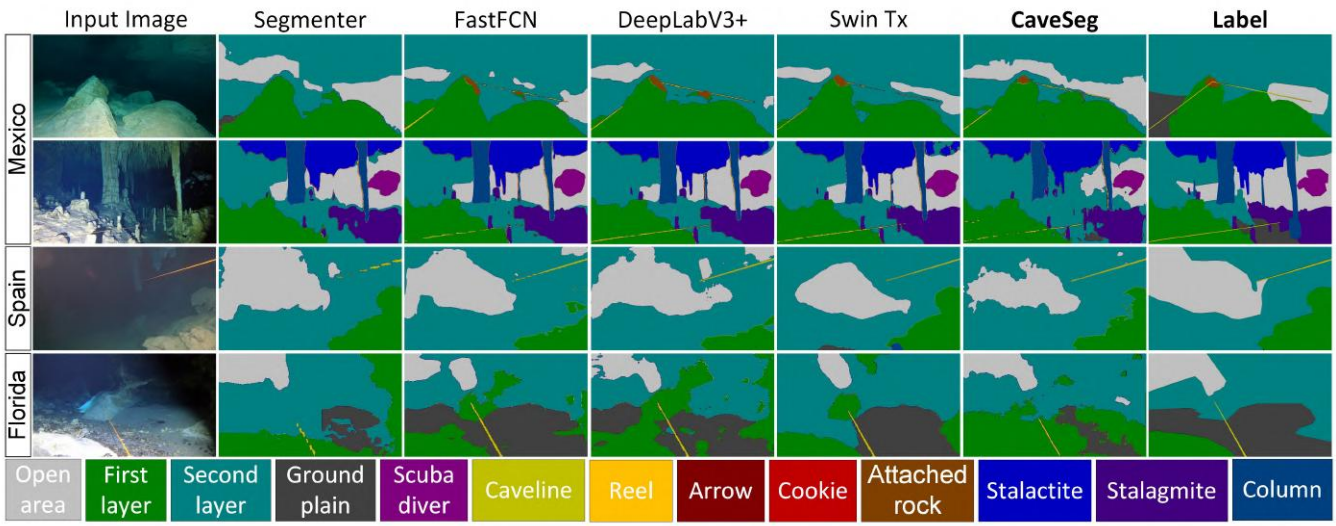

Fig. 5: A few qualitative performance comparisons of all models on CaveSeg-Challenge test set are shown (results for only four top-performing models are shown for clarity). Note that the object detection and localization accuracy for categories such as caveline, open area, and navigation markers are particularly important for AUV navigation.

TABLE II: Computational complexities for all models are compared based on number of parameters, memory requirements in Mega-Bytes (MB), and inference speeds in FPS. All experiments are performed on an Nvidia™ A100 GPU server with 16 GB RAM.

Method	# Params ↓	Memory ↓	Speed ↑
FastFCN	66 M	548.32 MB	16.89 FPS
DeepLabV3+	42 M	489.38 MB	15.04 FPS
Segmenter	98 M	798.04 MB	13.73 FPS
Segformer	82 M	956.62 MB	10.92 FPS
Swin Tx	120 M	1380.16 MB	12.39 FPS
CaveSeg	**35** M	**406.40** MB	**19.78** FPS

B. Qualitative Evaluation

Figure 5 compares the qualitative performance of CaveSeg with other SOTA models. A unique feature of CaveSeg and other transformer-based models is that they perform better in finding large and connected areas, *e.g.*, categories such as first/second layer obstacles and open areas. The continuous regions segmented by the CaveSeg model facilitate a better understanding of the surroundings compared to DeepLabV3+ and FastFCN. This validates our intuition that window-shifting technique can indeed extract global features more accurately. While the shifted window method focuses on global feature extraction, the deeper layers and multi-scale pooling of PPM help to preserve the details of local features. The precise detection of caveline, which is only a few pixels wide, further supports this design choice.

VI. Use Cases: Vision-based Cave Exploration & Semantic Mapping by AUVs

A. Safe AUV Navigation Inside Underwater Caves

The confined nature of the underwater cave environment makes obstacle avoidance a particularly important task. Safe navigation approaches, such as AquaNav proposed by Xanthidis *et al*. [47] and especially AquaVis [48] can generate smooth paths by avoiding obstacles. To this end, having a holistic semantic map of the scene can ensure that the AUV can safely follow the caveline and keep it in the FOV during cave exploration. Our previous work [6] demonstrates the utility of caveline detection and following for autonomous underwater cave exploration. While caveline detection is paramount, having semantic knowledge about the surrounding objects in the scene is essential to ensure safe AUV navigation. As shown in Fig. 6, cavelines can be obscure in particular areas due to occlusions and blends inside the cave passages. Hence, dense semantic information provided by CaveSeg is important to ensure continuous tracking and re-discovery of the caveline and other navigation markers. A few more challenging scenarios are shown Fig. 7, where simply following the caveline is not sufficient due to the cluttered scene geometry. With the semantic knowledge of first layer and second layer obstacles, caveline, and open areas – an AUV can plan its trajectory safely and efficiently.

Fig. 6: A BlueROV2 is being teleoperated through a cave passage in Orange Grove, FL where the caveline is almost invisible; note that the yellow wire on the top image is the ROV's tether, not a caveline. In such scenarios, an AUV can leverage the segmented obstacles and open areas in the scene (overlayed on the right column) toward planning a trajectory to rediscover and follow the caveline.

B. Working With or Alongside Human Scuba Divers

The underwater cave environment is extremely hostile due to the lack of direct access to the surface. Cave divers enforce strict protocols on light configurations (always a primary and two backup lights), use of breathing gas (use only 30%

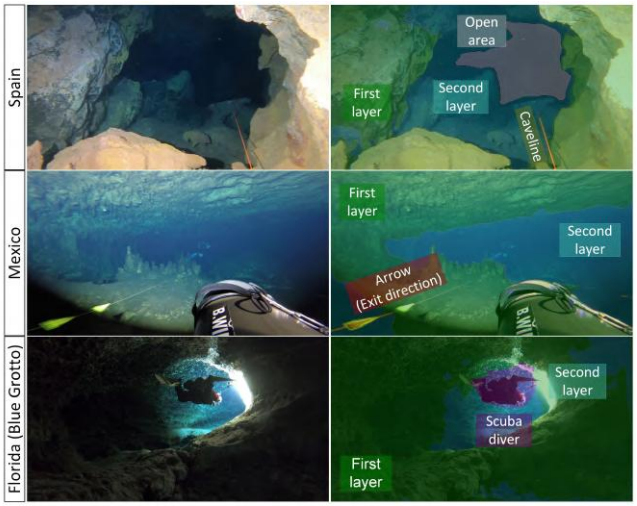

Fig. 7: A few examples of semantic scene parsing by CaveSeg for important navigation information are shown. The first scene shows a traceable caveline towards the open area and surrounding obstacles. When no open areas are visible (second image), arrows on the caveline are helpful to know the exit direction of the cave. Lastly, the accurate detection of human divers is a crucial step for giving the right-of-way as well as diver-robot cooperation.

on the way in, leaving 30% for the return and 30% for emergencies), and always keep near the caveline, ensuring there is an uninterrupted line to open water [5]. Of particular importance is the right of way for an exiting team. As such, CaveSeg maintains a label for divers (see Fig. 7) to ensure appropriate actions by the AUV. Specifically, in the presence of a diver, the lights of the vehicle will be dimmed so that the approaching diver can see the caveline. The AUV will reduce its speed, and refrain from using the downward-facing propellers in order to avoid stirring up the sediment; see Fig. 8 where the teleoperated ROV disturbed the sediment on the floor. Finally, the vehicle will move away from the caveline towards the walls of the cave.

Fig. 8: A BluROV2 is being teleoperated inside the Blue Grotto cave, FL: (left) entering the cave; (right) stirring the sediments.

C. 3D Semantic Mapping and State Estimation

Another prominent use case of CaveSeg is the 3D estimation and reconstruction of caveline for AUV navigation. Specifically, 2D caveline estimation can be combined with camera pose estimation from Visual-Inertial-Odometry (VIO) to generate 3D estimates. This could potentially be used to compare and improve manual surveys of existing caves. Moreover, 3D caveline estimations can also be utilized to reduce uncertainty in VIO and SLAM systems since the line directions in 3D point clouds can provide additional spatial constraints [49]. We demonstrate a sample result in Fig. 9, where *ray-plane triangulation* of cavelines is achieved using Visual-Inertial pose estimation for enabling 3D perception capabilities for underwater cave exploration. This experiment is performed on field data collected by ROVs in the Orange Grove cave system, FL, USA; we are currently exploring these capabilities for more comprehensive 3D semantic mapping of underwater caves.

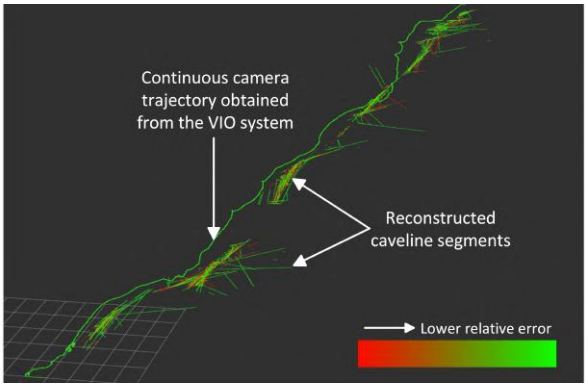

Fig. 9: Ray-plane triangulation of cavelines using VIO pose estimation and 2D line segments from deep segmentation pipeline are shown. The reconstructed 3D Lines are colored based on average re-projection error within its neighborhood in a connectivity graph.

VII. Conclusion & Future Work

This paper presents CaveSeg – the first comprehensive dataset and deep visual learning pipeline for underwater cave scene segmentation. With a focus on vision-guided cave exploration and mapping, we include semantic labels for object categories such as caveline, layered obstacles, arrows, cookies, attachment points, reels, and human scuba divers. We perform benchmark evaluations of SOTA models demonstrating the utility of such dataset for semantic segmentation of underwater cave scenes. We further propose a computationally light model that offers up to $1.8\times$ faster inference in addition to providing SOTA performance. We demonstrate that the predicted scene parsing labels can be utilized for safe AUV navigation inside underwater caves. We are currently investigating the scope of augmenting semantic maps with geometric information from a Visual-Inertial SLAM system such as SVIn2 [10]. This will potentially reveal the 3D position of the caveline as well as the relative location of obstacles and markers.

VIII. Acknowledgements

This research has been supported in part by the NSF grants 1943205 and 2024741. The authors would also like to acknowledge the help from the Woodville Karst Plain Project (WKPP), El Centro Investigador del Sistema Acuífero de Quintana Roo A.C. (CINDAQ), Global Underwater Explorers (GUE), Ricardo Constantino, and Project Baseline in collecting data and providing access to challenging underwater caves. We also appreciate Evan Kornacki for coordinating our field trials. Lastly, we thank Richard Feng, Ailani Morales, and Boxiao Yu for their assistance in collecting and annotating preliminary data in this project. The authors are also grateful for equipment support by Halcyon Dive Systems, Teledyne FLIR LLC, and KELDAN GmbH lights.

References

[1] A. H. G. González, C. R. Sandoval, A. T. Mata, M. B. Sanvicente, and E. Acevez, "The Arrival of Humans on the Yucatan Peninsula: Evidence from Submerged Caves in the State of Quintana Roo, Mexico," *Current Research in the Pleistocene*, vol. 25, pp. 1–24, 2008.

[2] K. Richmond, C. Flesher, N. Tanner, V. Siegel, and W. C. Stone, "Autonomous Exploration and 3-D Mapping of Underwater Caves with the Human-portable SUNFISH® AUV," in *Global Oceans 2020: Singapore–US Gulf Coast*, 2020, pp. 1–10.

[3] D. Ford and P. Williams, "Introduction to karst," in *Karst Hydrogeology and Geomorphology*. John Wiley & Sons, Ltd, 2007, ch. 1, pp. 1–8.

[4] P. L. Buzzacott, E. Zeigler, P. Denoble, and R. Vann, "American Cave Diving Fatalities 1969-2007," *International Journal of Aquatic Research and Education*, vol. 3, no. 2, p. 7, 2009.

[5] S. Exley, *Basic cave diving: A Blueprint for Survival*. Cave Diving Section of the National Speleological Society, 1986.

[6] B. Yu, R. Tibbetts, T. Barua, A. Morales, I. Rekleitis, and M. J. Islam, "Weakly Supervised Caveline Detection For AUV Navigation Inside Underwater Caves," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[7] J. Burge, *Underwater Cave Surveying*. Cave Diving Section of the National Speleological Society, 1988.

[8] B. Joshi, M. Xanthidis, M. Roznere, N. J. Burgdorfer, P. Mordohai, A. Q. Li, and I. Rekleitis, "Underwater exploration and mapping," in *IEEE OES AUV Symposium*, Singapore, Sep. 2022, pp. 1–7.

[9] B. Joshi, M. Xanthidis, S. Rahman, and I. Rekleitis, "High definition, inexpensive, underwater mapping," in *IEEE International Conference on Robotics and Automation (ICRA)*, Philadelphia, PA, USA, 2022, pp. 1113–1121.

[10] S. Rahman, A. Quattrini Li, and I. Rekleitis, "SVIn2: A Multi-sensor Fusion-based Underwater SLAM System," *International Journal of Robotics Research*, vol. 41, no. 11-12, pp. 1022–1042, Jul. 2022.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[12] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context Encoding for Semantic Segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," *arXiv preprint arXiv:1412.7062*, 2014.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[15] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[16] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for Semantic Segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.

[17] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[18] M. Mohammadi, S.-E. Huang, T. Barua, I. Rekleitis, M. J. Islam, and R. Zand, "Caveline detection at the edge for autonomous underwater cave exploration and mapping," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, Jacksonville, FL, USA, Dec. 2023.

[19] J. Fortin, S. Meacham, D. Rissolo, C. Le Maillot, and F. Devos, "Environmental Challenges, Technical Solutions and Standard Operating Procedures for Data Collection in Photogrammetric Studies toward a Unified Database of Objects And Features in Underwater Caves in Mexico," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 659–666, 2021.

[20] D. Rissolo, A. N. Blank, V. Petrovic, R. C. Arce, C. Jaskolski, P. L. Erreguerena, and J. C. Chatters, "Novel Application of 3D Documentation Techniques at a Submerged Late Pleistocene Cave Site in Quintana Roo, Mexico," in *Digital Heritage*, 2015, pp. 181–182.

[21] A. Mallios, P. Ridao, D. Ribas, M. Carreras, and R. Camilli, "Toward Autonomous Exploration in Confined Underwater Environments," *Journal of Field Robotics*, vol. 33, no. 7, pp. 994–1012, 2016.

[22] N. Weidner, S. Rahman, A. Quattrini Li, and I. Rekleitis, "Underwater cave mapping using stereo vision," in *IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, May 2017, pp. 5709–5715.

[23] N. Weidner, "Underwater Cave Mapping and Reconstruction Using Stereo Vision," M.S. thesis, Computer Science and Engineering Department, University of South Carolina, Columbia, SC, 2017.

[24] B. Joshi, S. Rahman, M. Kalaitzakis, *et al.*, "Experimental Comparison of Open Source Visual-Inertial-Based State Estimation Algorithms in the Underwater Domain," in *IEEE/RSJ International Conference on*

Intelligent Robots and Systems (IROS), Macau, Nov. 2019, pp. 7221–7227.

[25] S. Rahman, A. Quattrini Li, and I. Rekleitis, "An Underwater SLAM System using Sonar, Visual, Inertial, and Depth Sensor," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Macau, (IROS ICROS Best Application Paper Award. Finalist), Nov. 2019, pp. 1861–1868.

[26] S. Rahman, A. Quattrini Li, and I. Rekleitis, "Contour based reconstruction of underwater structures using sonar, visual, inertial, and depth sensor," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Macau, Nov. 2019, pp. 8048–8053.

[27] Q. Massone, S. Druon, Y. Breux, and J. Triboulet, "Contour-based Approach for 3D Mapping of Underwater Galleries," in *Global Oceans 2020: Singapore–US Gulf Coast*, IEEE, 2020, pp. 1–6.

[28] W. Wang, B. Joshi, N. Burgdorfer, K. Batsos, A. Quattrini Li, P. Mordohai, and I. Rekleitis, "Real-Time Dense 3D Mapping of Underwater Environments," in *IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023, pp. 5184–5191.

[29] I. Alonso, M. Yuval, G. Eyal, T. Treibitz, and A. C. Murillo, "CoralSeg: Learning Coral Segmentation from Sparse Annotations," *Journal of Field Robotics (JFR)*, vol. 36, no. 8, pp. 1456–1477, 2019.

[30] M. J. Islam, R. Wang, and J. Sattar, "SVAM: Saliency-guided Visual Attention Modeling by Autonomous Underwater Robots," in *Robotics: Science and Systems (RSS)*, NY, USA, 2022.

[31] M. J. Islam, C. Edge, Y. Xiao, P. Luo, M. Mehtaz, C. Morse, S. S. Enan, and J. Sattar, "Semantic Segmentation of Underwater Imagery: Dataset and Benchmark," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE/RSJ, 2020.

[32] M. J. Islam, P. Luo, and J. Sattar, "Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception," in *Robotics: Science and Systems (RSS)*, Jul. 2020.

[33] N. F. F. Alshdaifat, A. Z. Talib, and M. A. Osman, "Improved Deep Learning Framework for Fish Segmentation in Underwater Videos," *Ecological Informatics*, vol. 59, p. 101 121, 2020.

[34] O. Ulucan, D. Karakaya, and M. Turkan, "A Large-scale Dataset for Fish Segmentation and Classification," in *2020 Innovations in Intelligent Systems and Applications Conference*, IEEE, 2020, pp. 1–5.

[35] M. Modasshir, S. Rahman, O. Youngquist, and I. Rekleitis, "Coral Identification and Counting with an Autonomous Underwater Vehicle," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Kuala Lumpur, Malaysia, (Finalist of T. J. Tarn Best Paper in Robotics), Dec. 2018, pp. 524–529.

[36] M. Modasshir, S. Rahman, and I. Rekleitis, "Autonomous 3D Semantic Mapping of Coral Reefs," in *12th Conference on Field and Service Robotics (FSR)*, Tokyo, Japan, Aug. 2019, pp. 365–379.

[37] M. Waszak, A. Cardaillac, B. Elvesæter, F. Rødølen, and M. Ludvigsen, "Semantic Segmentation in Underwater Ship Inspections: Benchmark and Data Set," *IEEE Journal of Oceanic Engineering*, 2022.

[38] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[39] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "Fastfcn: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation," *arXiv preprint arXiv:1903.11816*, 2019.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[41] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[43] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified Perceptual Parsing for Scene Understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[45] M. Contributors, *MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark*, `https://github.com/open-mmlab/mmsegmentation`, 2020.

[46] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene Parsing Through ade20k Dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 633–641.

[47] M. Xanthidis, N. Karapetyan, H. Damron, S. Rahman, J. Johnson, A. O'Connell, J. O'Kane, and I. Rekleitis, "Navigation in the presence of obstacles for an agile autonomous underwater vehicle," in *IEEE International Conference on Robotics and Automation*, Paris, France, 2020, pp. 892–899.

[48] M. Xanthidis, M. Kalaitzakis, N. Karapetyan, J. Johnson, N. Vitzilaios, J. O'Kane, and I. Rekleitis, "Aquavis: A perception-aware autonomous navigation framework for underwater vehicles," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 5387–5394.

[49] J. Mo, M. J. Islam, and J. Sattar, "Fast Direct Stereo Visual SLAM," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 778–785, 2021.