2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 1-5, 2023. Detroit, USA

Weakly Supervised Caveline Detection For AUV Navigation Inside Underwater Caves

Boxiao Yu[1], Reagan Tibbetts[2], Titon Barua[2], Ailani Morales[1], Ioannis Rekleitis[2] and Md Jahidul Islam[1]

***Abstract*— Underwater caves are challenging environments that are crucial for water resource management, and for our understanding of hydro-geology and history. Mapping underwater caves is a time-consuming, labor-intensive, and hazardous operation. For autonomous cave mapping by underwater robots, the major challenge lies in vision-based estimation in the complete absence of ambient light, which results in constantly moving shadows due to the motion of the camera-light setup. Thus, detecting and following the *caveline* as navigation guidance is paramount for robots in autonomous cave mapping missions. In this paper, we present a computationally light caveline detection model based on a novel Vision Transformer (ViT)-based learning pipeline. We address the problem of scarce annotated training data by a weakly supervised formulation where the learning is reinforced through a series of noisy predictions from intermediate sub-optimal models. We validate the utility and effectiveness of such weak supervision for caveline detection and tracking in three different cave locations: USA, Mexico, and Spain. Experimental results demonstrate that our proposed model, *CL-ViT*, balances the robustness-efficiency trade-off, ensuring good generalization performance while offering 10+ FPS on single-board (Jetson TX2) devices.**

I. Introduction

Underwater caves 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 [1]. Moreover, underwater caves often present a pristine *capsule* preserved in time with major archaeological secrets [2]. Underwater cave exploration and mapping by human divers, however, is a tedious, labor-intensive, extremely dangerous operation even for highly skilled people [3]. Therefore, enabling Autonomous Underwater Vehicles (AUVs) and Remotely Operated Vehicles (ROVs) to enter, navigate, map, and finally exit an underwater cave is important to ensure the safety and efficacy of a mapping mission, as well as to potentially generate more accurate maps; Fig. 1 shows an ROV deployment scenario inside the Ballroom cavern at Ginnie Springs, Florida.

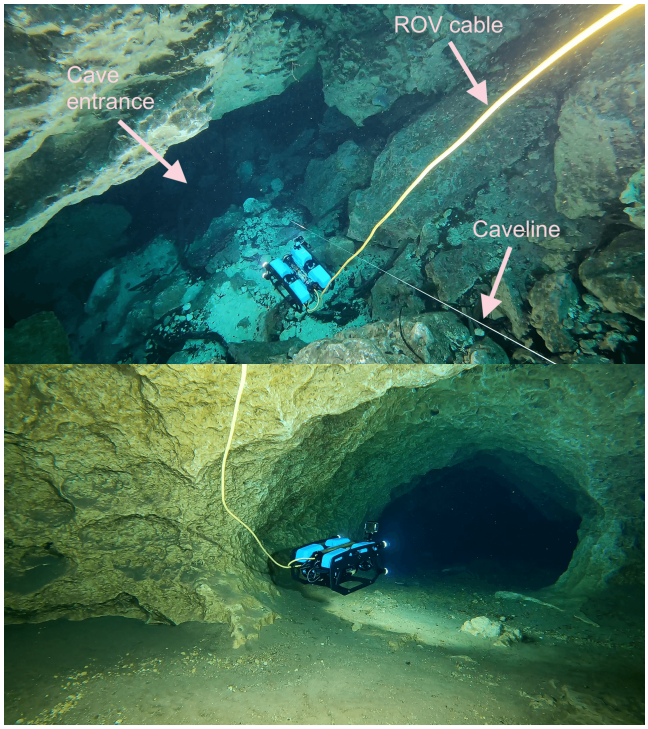


Fig. 1: A BlueROV2 operating inside the cave, Orange Grove Sink, Florida, USA. Note that the umbilical is connecting the ROV to a surface operator.

The first cardinal rule of cave diving as set by Sheck Exley is *"Always use a single, continuous guideline from the entrance of the cave throughout the dive."* [4]. Such guidelines, from here on termed *cavelines*, exist in all explored underwater caves and they provide the skeleton of the main passages. Mapping underwater caves is a multi-layered process. When a new section of a cave is discovered, a caveline is set identifying the passage. Consequently, the caveline is surveyed marking the depth and orientation at the points where the line is attached to the cave (floor, ceiling, or walls) and the distance between attachment points (called placements). These surveys produce a one-dimensional retraction of the three-dimensional environment. Recording all this information together with additional observations [5] such as distance to the walls, ceiling, and floor– is a challenging, time-consuming, and error-prone process.

Our earlier work [6] utilizing a GoPro-9 action camera resulted in high-precision camera trajectory estimation by a Visual-Inertial Odometry (VIO) algorithm [7], which is comparable to manually surveyed caveline (see Fig. 2). Moreover, the collected data are continuous spatiotemporal

[1]RoboPI Laboratory, Dept. of ECE, University of Florida, FL 32611, USA. Email: {boxiao.yu, ailanimorales}@ufl.edu, jahid@ece.ufl.edu.
[2]AFRL Laboratory, Dept. of CSE, University of South Carolina, Columbia, SC 29208, USA. {rbt,baruat}@email.sc.edu, yiannisr@cse.sc.edu.

This research has been supported in part by the National Science Foundation under grants 1943205, 1919647, 2024741, 2024541 and 2024653. The authors would also like to acknowledge the help of the Woodville Karst Plain Project (WKPP), El Centro Investigador del Sistema Acuífero de Quintana Roo A.C. (CINDAQ), and Ricardo Constantino, Project Baseline in collecting data, providing access to challenging underwater caves, and mentoring us in underwater cave exploration.

2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-9190-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/IROS55552.2023.10342435

videos, which we used to generate weakly labelled data, and demonstrated that iterative filtering of mislabelled samples can help regulate sample extraction for improved learning [8]. We found that the major challenges of data-driven solutions for problems such as the caveline detection and tracking, are: (i) learning from very few annotated samples; and (ii) ensuring generalization performance across different waterbodies, scene geometries, and optical degradations.

In this work, we address the aforementioned issues by developing a weakly supervised Vision Transformer (ViT)-based learning pipeline for autonomous caveline detection by AUVs. We demonstrate that with a limited amount of annotated training samples, learning can be reinforced iteratively from intermediate sub-optimal solutions. Specifically, after each *training phase*, the *weak* predictions are carefully sorted by a human expert into positive (*i.e.*, accurate) and negative (erroneous) labels. The noisy positive samples and a fraction of newly annotated negative samples are then fused to reinforce learning in the subsequent phases. With a series of experiments, we show that robust caveline detection with good generalization performance can be achieved with only 1.5K-2K annotated samples within 2-3 training phases.

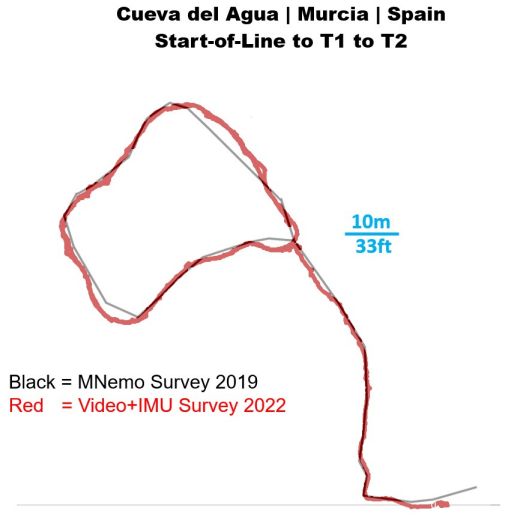


Fig. 2: Estimated trajectory together with manually measured ground truth from baseline; Cueva Del Agua, Spain.

We conduct experiments on **three cave systems** in different geographical locations: the Devil's system in Florida, USA; Dos Ojos Cenote, QR, Mexico; and Cueva del Agua in Murcia, Spain. We compile the data into three sets containing different types of cavelines, in terms of thickness and color, and different background and optical degradation levels. In order to ensure robustness, we evaluate both intra-set and inter-set detection performance – with the goal of achieving good generalization performance on data from an unseen location. We validated the effectiveness of our weakly supervised multi-phase training for several genres of prominent state-of-the-art (SOTA) models based on convolutional neural networks (CNNs), attention networks, conditional random fields (CRFs), U-shaped encoder-decoders, and ViTs.

Moreover, we develop a novel ViT-based learning pipeline named **CL-ViT** that offers the design choice of a *base model* with EfficientNetB5 [9] backbone and a *light model* with MobileNetV3 [10] backbone for offline use (by surface operators) and online processing (onboard AUVs), respectively. The base model surpasses SOTA detection performance and provides fine-grained caveline localization in image space. Additionally, with a highly efficient MobilenetV3 backbone [10] CL-ViT light model has only 12.67M parameters, which is about 51.55% less than DeepLabv3+ [11] and 46.61% less than PAN [12] – two of the best competitor baselines. As a result, it offers significantly faster inference rates: over 215.79 FPS on an NVIDIA™ RTX 3060 and 10.71 FPS on a single-board Jetson TX2. A series of challenging test experiments reveal that CL-ViT offers consistent performance for detecting cavelines with the presence of shadow, lighting variations, and other optical artifacts.

Furthermore, we developed a post-processing algorithm to filter the raw output masks of CL-ViT for more consistent and smooth caveline localization. We achieve this by first extracting a set of candidate lines from the binary output mask, then applying a voting procedure for non-maxima suppression based on the *accumulator space* of the probabilistic Hough transform [13]. We demonstrate the effectiveness of this post-processing step qualitatively for noisy predictions in various challenging scenarios as well.

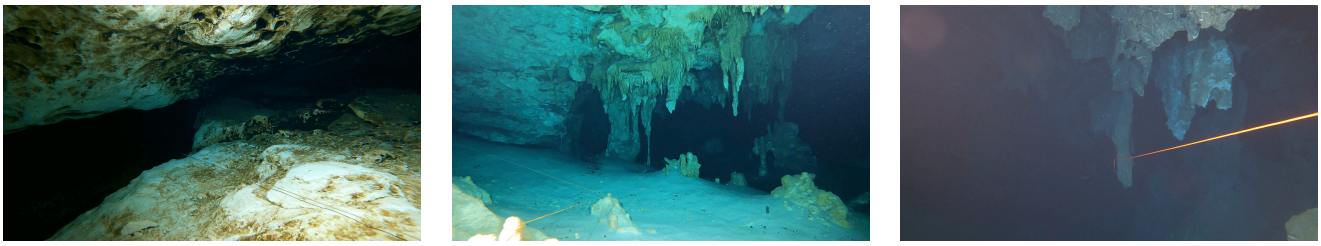
II. Background & Related Work

A. Underwater Cave Mapping and Exploration

In order to produce informative representations of underwater caves, divers typically use photogrammetry [14], with a special focus on recording archaeological sites [2], [15]. Attempts to automate underwater cave mapping by AUVs have proven to be challenging and thus remain an open problem. Mallios *et al*. [16] deployed an AUV to manually collect acoustic data from inside a cave for offline mapping, whereas Weidner *et al*. [17], [18] utilized a stereo camera to map the walls of a cave. It is worth noting that vision-based underwater state estimation is extremely challenging due to the lighting variations, light absorption, and blurriness [19]. More recently, Rahman *et al*. [7], [20], [21] presented a framework where acoustic, visual, inertial, and water depth data are used to estimate the trajectory of the robot and also a sparse representation of the cave. Denser representation of the cave boundaries can be obtained by mapping the contours [22], the moving shadows [23] or via dense stereo reconstruction [24]. These approaches will be utilized to enhance the mapping of an AUV following the caveline safely in and out of the cave. Sunfish [25] a new man-portable AUV is currently being deployed in caves in Florida.

B. Object Detection/Segmentation in Underwater Imagery

An essential capability of visually guided AUVs is to identify relevant objects and interesting image regions to make effective navigational decisions in real time. Various model-based techniques are generally deployed in fast visual search [26], [27], enhanced object detection [28], [29], and



(a) Devil system, FL, USA (b) Dos Ojos Cenote, QR, Mexico (c) Cueva del Agua, Murcia, Spain

Fig. 3: Datasets used in our experiments are collected using a GoPro-9 camera at three different locations. A sample for each dataset is shown; notice the (a) grey/white thinner cavelines in the *Florida dataset*; (b) thick yellow cavelines and a decorated background in the *Mexico dataset*; and (c) thick orange cavelines in the *Spain dataset*.

monitoring applications [30], [31]. For instance, Koreitem *et al*. [26] used a bank of pre-specified image patches to learn a similarity operator that guides the robot's visual search in an unconstrained setting. Besides, model-free approaches are more feasible for autonomous exploratory applications [32]. For instance, Girdhar *et al*. [33] formulated an online topic-modeling scheme that encodes visible features into a low-dimensional semantic descriptor for AUV exploration. More recent work by Modasshir *et al*. combined a deep learning-based classifier model with VIO to identify and track the locations of different types of corals to generate semantic maps [34] as well as volumetric models [35].

Due to the difficulties in acquiring large-scale labeled underwater data, the existing systems attempt to collect and annotate small-scale application-specific image data [36], [37]. Islam *et al*. [38] considered eight object categories for human-robot cooperative missions: robots, human divers, wrecks/ruins, aquatic plants, fish, reefs, and sea floor. Other datasets consider even fewer object categories such as marine debris or ship hull defects [39]. With limited training samples per object category over only a few waterbody types, it is extremely challenging to achieve good generalization performance by deep learning-based models for image recognition tasks. These limitations call for learning adaptations [40] with limited supervision and rigorous model design to ensure robust underwater visual perception.

III. Weakly Supervised Caveline Detection

A. Problem Formulation and Data Preparation

We formulate the problem of caveline detection in the RGB space as a binary image segmentation task, *i.e*., identifying pixels with caveline as a semantic map [41]. In our task, the background pixels and caveline pixels are assigned with 0 and 1 labels, respectively. For data-driven training and evaluation, we extract video frames from our cave exploration experiments [6], [7] conducted in three different locations: the Devil's system in Florida, USA; the Dos Ojos Cenote, QR, Mexico; and the Cueva del Agua in Murcia, Spain. We grouped the caveline frames from these locations into three datasets, which we term as the *Florida*, *Mexico*, and *Spain* dataset, respectively.

As illustrated in Fig. 3, we found that the three cave locations exhibit different caveline characteristics in terms of thickness, color, and background patterns. The cavelines in Florida are thin and off-white colored, whereas the Mexico caves are the most decorated with yellow colored lines. Cavelines in Spain are also thick and of orange color. In general, the main cavelines in popular locations are thicker, while off the main path become thinner; the grey/white colored lines take a darker color over time and often blend with the background patterns. We identify these variety of challenging cases and prepare 1050 images in each set, totaling $3 \times 1050 = 3150$ instances. We focused on maximizing variance in the data by including varieties in caveline color, distance, background/waterbody patterns as well as different cave formations (*e.g*., stalactites, stalagmites, columns) and navigational aids such as arrows and cookies. Four human participants sorted these image samples and then pixel-annotated the cavelines for ground truth generation, which we utilizes for the training and evaluation of all models.

B. Proposed Model: ***CL-ViT***

We develop a lightweight caveline detection model CL-ViT for use by visually guided AUVs in underwater cave mapping and exploration tasks. To this end, we focus on enabling two important features: (i) robustness to noisy low-resolution inputs because cavelines are only a few pixels wide even in a high-resolution camera feed; and (ii) efficient inference on single-board embedded platforms. We attempt to achieve this in CL-ViT model by integrating multi-scale local hierarchical features and global spatial information for efficient pixel-wise segmentation of cavelines. CL-ViT consists of two major learning components: an efficient encoder-decoder backbone and a ViT-based refinement module; the network architecture is illustrated in Fig. 4.

1) Choice of Backbones: We incorporate two options for the deep hierarchical feature extraction in CL-ViT: (i) a light model with MobileNetV3 [10] backbone for on-board AUV processing; and (ii) a base model with EfficientNetB5 [9] backbone for offline use, *e.g*., when human operators on surface control ROVs inside a cave. The MobileNetV3 is a lightweight CNN-based model designed for resource-constrained platforms. The encoder contains a series of fully convolutional layers with 16 filters followed by 15 residual bottleneck layers. We then use a mirrored decoder with six convolutional blocks to map the encoded features into 48 filters of 480×270 resolution (with an input of $960 \times 540 \times 3$). On the other hand, EfficientNet uses a technique called *compound coefficient* to scale up models in a simple but effective manner. It uniformly scales features in width, depth,

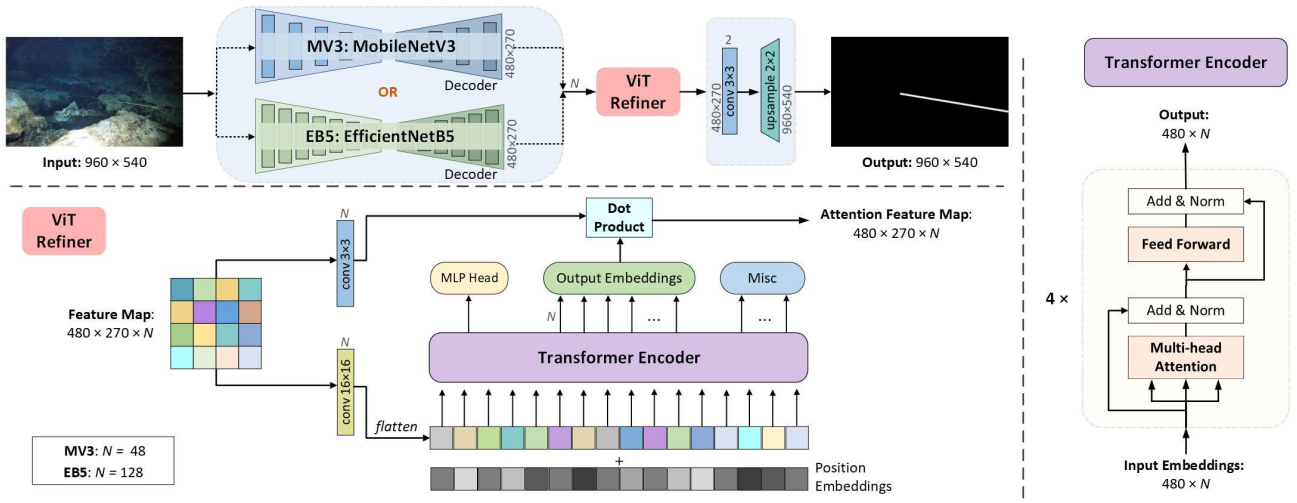


Fig. 4: The end-to-end learning pipeline of our proposed CL-ViT model is shown. Input images are first fed to the backbone (light model: MobileNetV3 backbone; base model: EfficientNetB5 backbone) for feature extraction. Those features are forwarded to our transformer-based refinement module (ViT Refiner), followed by a convolution and upsampling block to generate the caveline detection mask.

and resolution to ensure effective receptive fields for feature extraction. We use EfficientNetB5 which extracts 128 filters of 480×270 resolution from $960 \times 540 \times 3$ inputs.

2) ViT-based Refinement Module: Following feature extraction, we design a ViT-based refinement module to *transform* and *embed* the contextual features into an efficient prediction head. Our idea is to allow each feature position to have consistent receptive fields so that the global spatial information is accurately embedded. As shown in Fig. 4, the N=48/128 filters extracted by the backbone are 16×16 convolved and flattened to *patch embeddings*, which are concatenated with learnable *position embeddings*. These embeddings are then propagated to the transformer encoder, which applies a four-layer multi-head attention mapping. Subsequently, the normalized and 3×3 convolved feature maps are projected (dot product operation) with N output embeddings; the remaining embeddings in the MLP head are dropped. The selected attention maps then generate the binary caveline segmentation map after a final convolution and upsampling operation at 960×540 resolution.

3) Learning Objective: The end-to-end training is driven by two loss functions: the standard cross-entropy loss [42] and the Dice loss proposed by Milletari *et al*. [43]. The cross-entropy loss quantifies the dissimilarity in pixel intensity distributions between the generated caveline map ($\hat{y}$) and its ground truth (y). For a total of n_p pixels, it is calculated as:

$$\mathcal{L}_{BCE} = \frac{1}{n_p} \sum\nolimits_i [-y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)]. \quad (1)$$

While we initially trained all caveline detectors with $\mathcal{L}_{BCE}$ alone, we noticed a severe class imbalance problem, since there are very few positive (caveline) pixels compared to the negative (background) pixels. We address this issue by adding the Dice loss, which balances foreground and background classes by normalizing y and $\hat{y}$ as follows:

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_i y_i \hat{y}_i}{\sum_i y_i^2 + \sum_i \hat{y}_i^2}. \quad (2)$$

Finally, the end-to-end learning objective is formulated as:

$$\mathcal{L}_{CL} = \lambda_{CE}\, \mathcal{L}_{BCE} + \lambda_D\, \mathcal{L}_{Dice}. \quad (3)$$

Here, we find the λ_{CE} and λ_D empirically for different models independently through hyper-parameters tuning.

4) Weakly-Supervised Iterative Training: Pixel-annotated training data is very scarce for unique problems such as caveline detection in underwater caves. As discussed earlier, caveline characteristics and background waterbody patterns in each cave locations differ greatly, making it difficult to compile a comprehensive dataset for supervised training. We address this limitation by a weakly supervised formulation, where model accuracy is improved incrementally on new locations' data. This speeds up model adaptations during robotics field deployments to a new location by eliminating the need for labeling entire datasets for supervised training. We validate this hypothesis on each dataset (*i.e*., Florida, Spain, and Mexico data) based on the *leave-one-out* mechanism, as illustrated in Fig. 5.

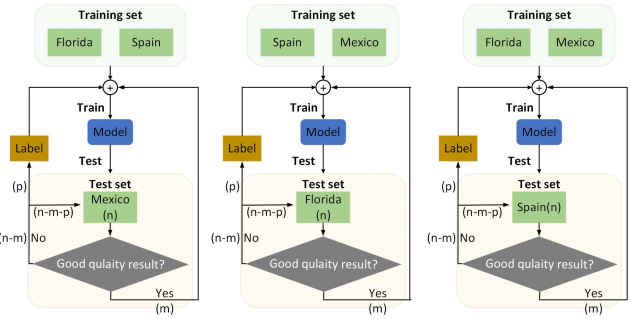


Fig. 5: Our weakly-supervised iterative training process is shown.

For each of the three cases shown in Fig. 5, the weak supervision is carried out as follows. The initial model is evaluated on the full test set (of 1050 samples), from which a human expert sorts out the good quality predictions to reinforce the learning in the next phase. The human expert also selects a set of challenging samples where the

TABLE I: Quantitative comparison for caveline detection performance by CL-ViT and other SOTA models are shown for our weakly supervised iterative learning phases (see Sec. III-B.4 for the discussion on how these training phases are carried out).

Training data	**Test set**	**Phase**	**Metric** (↑)	EMANet	UNet	DPT	PAN	DeepLabv3+	**CL-ViT** (MV3)	**CL-ViT** (EB5)
Florida + Spain	Mexico	1	IoU	13.43	32.65	35.07	38.39	46.16	25.08	**50.92**
			F1	77.39	66.35	68.42	82.12	84.65	59.04	**85.72**
		2	IoU	15.90	45.56	51.21	62.41	61.90	33.08	**68.73**
			F1	83.35	82.50	89.29	96.62	95.99	73.95	**97.84**
		3	IoU	15.98	46.05	54.07	55.84	60.57	32.60	**70.28**
			F1	79.74	83.45	88.53	95.04	96.98	72.45	**98.34**
Florida + Mexico	Spain	1	IoU	13.10	55.32	48.23	54.86	58.19	42.07	**66.47**
			F1	77.67	91.24	84.60	93.65	95.43	84.24	**96.03**
		2	IoU	24.45	60.30	64.23	66.78	70.70	48.78	**76.00**
			F1	92.54	96.80	97.14	98.07	98.47	92.92	**98.48**
		3	IoU	18.67	62.88	67.24	69.86	71.13	48.29	**77.39**
			F1	79.13	96.93	97.06	98.39	98.64	91.71	**98.74**
Spain + Mexico	Florida	1	IoU	6.87	19.21	18.89	27.87	28.78	16.18	**39.11**
			F1	31.45	37.50	35.45	55.90	56.50	37.68	**74.23**
		2	IoU	13.60	24.24	26.17	28.53	33.30	20.64	**41.16**
			F1	65.66	52.75	57.24	71.64	77.73	54.49	**85.26**
		3	IoU	13.20	26.49	22.01	34.81	31.67	15.79	**40.53**
			F1	69.33	57.79	51.23	79.21	76.71	47.69	**85.00**

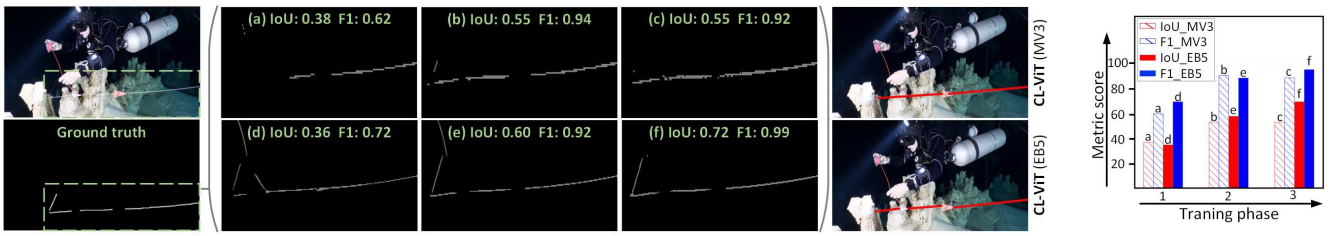


Fig. 6: Effectiveness of our weakly supervised caveline detection pipeline is shown with an example. Output maps a/d, b/e, and c/f are the test results for CL-ViT model with MobileNetV3/EfficientNetB5 backbone after the first, second, and third phase of training, respectively. As seen, the visual prediction results and metric scores gradually get better after each learning phase. The final post-processed predictions are also shown on the right overlayed on the input image.

model failed, then annotates and combines them into the training set in order to balance distribution positive (accurate) and negative (erroneous) samples. This process is repeated several times until a satisfactory number of training samples are compiled. Our experiments reveal that we get 15%-20% good quality predictions in the first phase and another 34%-47% in the second phase. All 1050 images get labelled within 3 phases, where human experts relabelled only 200-250 images as negative samples. Thus the remaining labels are *weak labels* generated by intermediate sub-optimal models for subsequent weak supervision.

5) Post-processing: In the post-processing step, we smooth the raw CL-ViT output of binary pixel predictions into continuous line segments. We achieve this by first interpolating a sequence of connected straight line segments by modeling a probabilistic Hough transform [13]. Then we apply a voting procedure for non-maxima suppression and generate the most dominant line. To ensure robustness of this suppression mechanism for all types of noisy and incomplete predictions, we empirically tuned the hyper-parameters, *e.g.*, the distance metric, acute angle threshold for merging pair-wise lines, and the number of iterations.

IV. Experimental Results

A. Baseline Models and Evaluation metrics

We developed a unified training pipeline for SOTA models across the CNN, CRF, and ViT literature. Specifically, we use: EMANet [44], UNet [45], DPT [46]. PAN [12], and DeepLabv3+ [11] for baseline performance analyses. We use Pytorch libraries to implement a unified learning pipelines for CL-ViT and all SOTA models in comparison. RMSprop [47] is used as the optimizer with an initial learning rate of 10^{-5}, a momentum of 0.9, and a weight decay of 10^{-8}. The input-output resolution is set to 960×540 for all models; other SOTA model-specific parameters are chosen based on their respective recommended configurations.

For performance evaluation, we use two standard metrics: **IOU** and **F1 score**. The IoU (Intersection Over Union) measures caveline localization performance using the area of overlapping regions of the predicted and ground truth labels. it is defined as $IoU = \frac{Area\ of\ overlap}{Area\ of\ union}$. Besides, the F1 score quantifies the correctness of predicted labels compared to ground truth by the normalized precision ($\mathcal{P}$) and recall ($\mathcal{R}$) scores as $\mathcal{F} = \frac{2 \times \mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}}$.

B. Qualitative and Quantitative Evaluation

1) Effectiveness of Weak Supervision: We first demonstrate the utility and effectiveness of our weakly supervised learning pipeline for the three cases depicted in Fig. 5. The corresponding quantitative results are listed in Table I, which shows that all models exhibit incremental improvements over learning phases 1, 2, and 3. This validates our intuition that robust generalization performance by standard deep visual learning models can be achieved with very few labeled data from a new location for fast model adaptation. Fig. 6

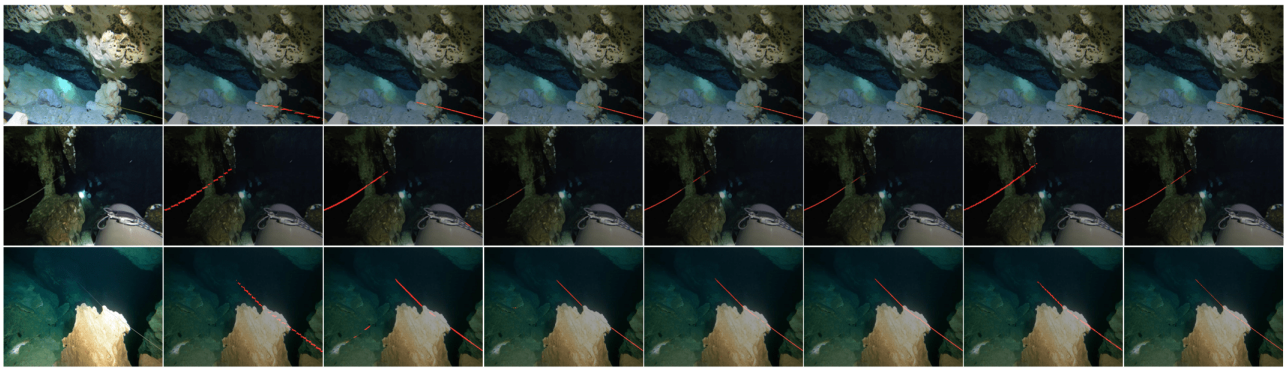


Fig. 7: A few qualitative comparisons are shown for caveline detection by CL-ViT and other SOTA models on cross-location test images. The DeepLabv3+, DPT, and PAN models provide well-localized predictions, while CL-ViT (EfficientNetB5) generates the most fine-grained caveline detection (*i.e.*, thinnest continuous lines). Note that all images are overlayed with raw outputs without post-processing.

shows a particular example where IoU scores improved from 0.36/0.38 to 0.55/0.72, while F1 scores improved from 0.62/0.72 to 0.92/0.99 for the CL-ViT light/base model, respectively. The generated maps become increasingly fine-grained as well; the final output maps can be further post-processed for well-localized detection of cavelines.

2) Performance Analyses of CL-ViT: We conduct a thorough performance evaluation of CL-ViT and other SOTA models based on all cross-location test images. A few qualitative comparisons are shown in Fig. 7, which shows consistent results from CL-ViT, DeepLabv3+, DPT, and PAN. Our CL-ViT (EfficentNetB5) model achieves the most fine-grained caveline detection performance with the thinnest continuous line segments. While not as fine-grained, oue light CL-ViT (MobileNetV3) model localizes the cavelines reasonably as well, which can be further refined by post-processing.

We show the corresponding quantitative test results in Table I; it confirms the superior performance from CL-ViT (EfficentNetB5) for both IoU and F1 score metrics. Although CL-ViT (MobileNetV3) does not surpass the SOTA performance, with a significantly lighter model architecture, it offers 51.52%-89.74% memory efficiency and 7-43 times faster inference rates. It runs at **10+ FPS rates on Nvidia™ Jetson TX2** devices, which makes it feasible for single-board deployments in AUVs' autonomy pipeline.

TABLE II: Quantitative test results are shown for CL-ViT and other **top three** models from Table I; their memory requirements in Mega-Bytes (MB), and inference rates in FPS (Core i9-12900 CPU) and FPS* (single-board Jetson TX2) are compared as well.

Metric	UNet	DPT	DeepLabv3+	**CL-ViT** (MV3)	**CL-ViT** (EB5)
↑ IoU	38.34	38.88	49.77	28.57	**58.30**
↑ F1	86.68	77.89	93.07	77.79	**95.87**
↑ FPS	2.34	0.46	2.41	**20.21**	0.77
↑ FPS*	1.15	0.23	1.19	**10.71**	0.38
↓ MB	124.20	496.20	105.00	**50.90**	313.70

3) Challenging Cases: As discussed earlier, very low resolutions of positive (caveline) pixels compared to the negative (background) pixels cause the *class imbalance* problem in caveline detection learning. While we eliminate this by using a relatively high input resolution of 960×540, there

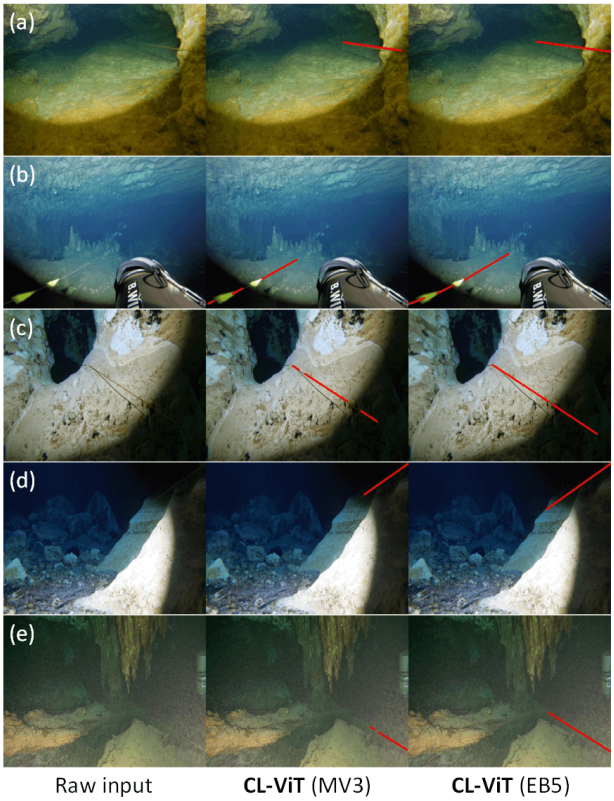


Fig. 8: Post-processed CL-ViT output for a few test cases from the **CL-Challenge** dataset; notice the (a) lack of contrast in the caveline regions; (b) presence of arrows/cookies; (c) caveline shadow appears similar to another line; (d) caveline is outside the illuminated area; (e) scattering and distortions around the caveline.

are other scenarios where CL-ViT (and other models) are faced with challenges. We identify a subset of such cases and compile a **CL-Challenge** test set with 200 samples. It includes images with severe optical distortions, lack of contrast, over-saturation, shadows, low-light conditions, occlusion, and other issues that make it extremely challenging to locate the caveline, even for a human observer. As shown in Fig. 8, CL-ViT models are still able to localize the caveline for the most part. Despite some noisy predictions, these inspiring results indicate that caveline detection by CL-ViT can facilitate safe AUV navigation inside underwater caves.

V. Conclusions

In this paper, we presented a novel learning pipeline for fast caveline detection in images from underwater caves. We formulated a weakly supervised approach that facilitates a rapid model adaptation to data from new location by requiring very few ground truth labels. A comparison with SOTA frameworks demonstrated higher accuracy and efficiency of the proposed approach. Cavelines traverse the majority of explored underwater caves providing a roadmap that can guide a robot inside a cave and then safely back out. Of paramount importance is robustness across different appearances both of the line but also of the surrounding background. Tests on three different locales demonstrated accurate performance across different domains and lines. Currently, we are developing an autonomous caveline-following system that uses CL-ViT's caveline predictions in tandem with a VIO system for visual servoing.

References

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

[2] 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.

[3] 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.

[4] S. Exley, *Basic cave diving: A blueprint for survival*. Cave Diving Section of the National Speleological Society, 1986.

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

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

[7] 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.

[8] M. Modasshir and I. Rekleitis, "Augmenting coral reef monitoring with an enhanced detection system," in *IEEE International Conference on Robotics and Automation*, Paris, France, 2020, pp. 1874–1880.

[9] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019, pp. 6105–6114.

[10] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for mobilenetv3," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324.

[11] 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.

[12] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[13] N. Kiryati, Y. Eldar, and A. M. Bruckstein, "A probabilistic hough transform," *Pattern recognition*, vol. 24, no. 4, pp. 303–316, 1991.

[14] 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.

[15] 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.

[16] 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.

[17] 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)*, 2017, pp. 5709–5715.

[18] 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.

[19] 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)*, 2019, pp. 7221–7227.

[20] S. Rahman, A. Quattrini Li, and I. Rekleitis, "Sonar Visual Inertial SLAM of Underwater Structures," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 5190–5196.

[21] 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)*, 2019, pp. 1861–1868.

[22] 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.

[23] 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)*, 2019, pp. 8048–8053.

[24] 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.

[25] 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.

[26] K. Koreitem, F. Shkurti, T. Manderson, W.-D. Chang, J. C. G. Higuera, and G. Dudek, "One-shot informed robotic visual search in the wild," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5800–5807.

[27] 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.

[28] J. Zhu, S. Yu, L. Gao, Z. Han, and Y. Tang, "Saliency-Based Diver Target Detection and Localization Method," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[29] M. J. Islam, Y. Xia, and J. Sattar, "Fast Underwater Image Enhancement for Improved Visual Perception," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 2, pp. 3227–3234, 2020.

[30] M. Modasshir, A. Quattrini Li, and I. Rekleitis, "Deep neural networks: A comparison on different computing platforms," in *Canadian Conference on Computer and Robot Vision (CRV)*, 2018, pp. 383–389.

[31] T. Manderson, J. C. G. Higuera, R. Cheng, and G. Dudek, "Vision-based Autonomous Underwater Swimming in Dense Coral for Combined Collision Avoidance and Target Selection," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1885–1891.

[32] Y. Girdhar and G. Dudek, "Modeling Curiosity in a Mobile Robot for Long-term Autonomous Exploration and Monitoring," *Autonomous Robots*, vol. 40, no. 7, pp. 1267–1278, 2016.

[33] Y. Girdhar, P. Giguere, and G. Dudek, "Autonomous Adaptive Exploration using Realtime Online Spatiotemporal Topic Modeling," *International Journal of Robotics Research (IJRR)*, pp. 645–657, 2014.

[34] 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)*, Dec. 2018, pp. 524–529.

[35] 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.

[36] 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.

[37] M. Ravanbakhsh, M. R. Shortis, F. Shafait, A. Mian, E. S. Harvey, and J. W. Seager, "Automated Fish Detection in Underwater Images Using Shape-Based Level Sets," *Photogrammetric Record*, vol. 30, no. 149, pp. 46–62, 2015.

[38] 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)*, 2020.

[39] 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.

[40] B. Yu, J. Wu, and M. J. Islam, "Udepth: Fast monocular depth estimation for visually-guided underwater robots," in *Accepted at the IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023.

[41] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: State of the art," *International journal of multimedia information retrieval*, vol. 9, no. 3, pp. 171–189, 2020.

[42] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[43] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D vision (3DV)*, 2016, pp. 565–571.

[44] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *IEEE/CVF Int. Conference on Computer Vision*, 2019, pp. 9167–9176.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2015, pp. 234–241.

[46] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *IEEE/CVF Int. Conference on Computer Vision*, 2021, pp. 12 179–12 188.

[47] T. Tieleman, G. Hinton, *et al.*, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.