

Knowledge Distillation for Feature Extraction in Underwater VSLAM

Jinghe Yang¹, Mingming Gong², Girish Nair¹, Jung Hoon Lee³, Jason Monty³, Ye Pu¹

Abstract—In recent years, learning-based feature detection and matching have outperformed manually-designed methods in in-air cases. However, it is challenging to learn the features in the underwater scenario due to the absence of annotated underwater datasets. This paper proposes a cross-modal knowledge distillation framework for training an underwater feature detection and matching network (UFEN). In particular, we use in-air RGBD data to generate synthetic underwater images based on a physical underwater imaging formation model and employ these as the medium to distil knowledge from a teacher model SuperPoint pretrained on in-air images. We embed UFEN into the ORB-SLAM3 framework to replace the ORB feature by introducing an additional binarization layer. To test the effectiveness of our method, we built a new underwater dataset with groundtruth measurements named EASI (<https://github.com/Jinghe-mel/UFEN-SLAM>), recorded in an indoor water tank for different turbidity levels. The experimental results on the existing dataset and our new dataset demonstrate the effectiveness of our method.

I. INTRODUCTION

At present, Autonomous Underwater Vehicles (AUVs) are competent for many tasks, such as repairing underwater facilities, natural gas extraction, and environmental safety inspections [1]. In order to complete complex control tasks, AUVs need to be able to locate themselves. However, limited sensor choices and poor visualization make localization challenging in the underwater environment. For example, the Global Positioning System (GPS) and high-frequency radio signals are unavailable in underwater environments. Consequently, AUVs must rely on other sensors to obtain their position information, such as cameras and sonars [2].

Visual SLAM (VSLAM) is one potential vision-based solution to localization and environmental perception in AUVs. VSLAM algorithms use cameras as the main exteroceptive sensor. The front end of VSLAM, named Visual Odometry (VO), accepts the perceived images and finds corresponding features in adjacent frames to estimate the current states [3]. VSLAMs can be divided into feature-based and direct methods depending on how the geometric constraints are extracted at the front end [4]. Feature-based VSLAM uses sparse feature points to construct 3D landmarks in space, while the direct methods estimate the motion trajectory based on the intensity of pixel values on continuous images. VSLAM is potentially a good candidate for underwater scenarios when AUVs are close to the bottom, and there is enough visibility to observe some objects or details on the

seabed. However, the performance of VSLAM is affected by the underwater environment. Joshi et al. [5] compared VSLAM algorithms on several underwater sequences. Their results show that the feature-based indirect method ORB-SLAM2 [6] performs more robust than other methods, e.g., LSD-SLAM [7] and DSO [8], but still suffers from the unstable performance of the feature extraction and matching due to the blurriness and low contrast of the underwater images.

For feature-based VSLAM such as ORB-SLAM2, the quality of the detected features and feature matching between neighbouring frames is a core factor of the performance. In [9], the experimental results show that attenuation and backscattering in underwater imaging reduce the effectiveness of the traditional handcrafted feature detection methods. Recently, learning-based feature extraction and matching have shown great potential for in-air scenarios [10] [11] [12]. For example, SuperPoint [12] is a self-supervised solution with a good performance on the HPatches dataset [13] compared with the handcrafted methods SIFT [14] and ORB [15]. However, when applying the learning-based methods to underwater environments, we face new challenges, for example, the absence of suitable underwater datasets for training and the blurriness of the underwater images, making it particularly challenging to define and label the actual feature points.

To tackle these challenges, we leverage the Cross-Modal Distillation [16] framework to transfer knowledge from in-air images to underwater images. More specifically, we distil knowledge from a teacher SuperPoint model trained on in-air images to our student underwater feature detection and matching network. To enable efficient knowledge transfer, we propose synthesising underwater images from in-air images using a physical model and designing proper loss functions to train the student network on synthetic underwater images. This framework transfers knowledge from a rich labelled modality as supervision to an unlabeled paired modality. Below is the list of our contributions in this paper:

- We propose the Underwater Feature Extraction Network (UFEN) - an underwater feature point detection and description neural network that directly extracts and matches feature points on underwater images.
- We use the idea of Cross-Modal Knowledge Distillation in the training of UFEN. We transfer the knowledge of the pre-trained model (SuperPoint) for the in-air scenarios to the UFEN model via synthetic underwater images.
- We propose a monocular underwater VSLAM sys-

¹The authors are with the Department of Electrical and Electronic Engineering, The University of Melbourne. ²The authors are with the School of Mathematics and Statistics, The University of Melbourne. ³The authors are with the Department of Mechanical Engineering, The University of Melbourne. Correspondence: jinghey@student.unimelb.edu.au

tem UFEN-SLAM¹, in which the UFEN is integrated into the state-of-art feature-based VSLAM framework - ORB-SLAM3 [17] by introducing an additional binarization layer to the end of the descriptor decoder.

- We built an underwater dataset referred to as EASI with the groundtruth measured by a motion tracking system with different water turbidity levels and evaluated the performance of our method on this dataset.

II. PRELIMINARIES

This section introduces the three main tools, which will be used in our algorithm, including a state-of-art VSLAM framework - ORBSLAM3 [17], a pre-trained neural network for feature extraction and matching - Superpoint [12], and an underwater image formation physical model for synthetic data generation.

A. ORB-SLAM3

A state-of-art feature-based method, ORB-SLAM3 consists of four components, including Atlas for representing the set of disconnected maps and three threads: i) tracking; ii) local mapping; iii) loop and map merging. It uses ORB features [15] and the image pyramid to identify spatial feature points at different distances. ORB feature detects corner points by the differences in pixel intensities and uses a 256-dimensional binary descriptor with orientation invariance. ORB is designed for in-air cases. Due to the attenuation of underwater sequences, the sharpness of the image is degraded, and the backscattering caused by particles can give the image different characteristics from in-air images. The experimental results in [5] show that on underwater sequences, the monocular ORB-SLAM2 [6] frequently fails to track because of the unstable performance of the ORB feature. Furthermore, it is hard to initialize if the image is blurry and in low contrast. However, it is still a better candidate than the direct methods, DSO and LSD-SLAM.

In Section IV, we will integrate the newly trained UFEN model into the ORB-SLAM3 framework for detecting feature points and assigning descriptors.

B. SuperPoint

SuperPoint [12] is a learning-based feature point extracting and matching neural network. As shown in Fig. 1, SuperPoint has one shared encoder and two decoders, including an interest point detector decoder and a descriptor decoder. SuperPoint takes grayscale images $U \in \mathbb{R}^{H \times W \times 1}$ of size $H \times W$ as its input. The interest point detector head computes $\mathcal{X} \in \mathbb{R}^{H_c \times W_c \times 65}$ and outputs a probability tensor $\in \mathbb{R}^{H \times W \times 1}$ after the channel-wise *Softmax* and *Reshape* operations. H_c and W_c are 1/8 of the input images size. The descriptor decoder head computes $\mathcal{D} \in \mathbb{R}^{H_c \times W_c \times 256}$ and finally outputs a 256-dimensional descriptor tensor $\in \mathbb{R}^{H \times W \times 256}$ after the bi-cubic interpolation and L2-normalization.

SuperPoint is initially trained on corner points of the synthetic general shapes. This guarantees the robust performance of SuperPoint when applied to different scenarios, tending

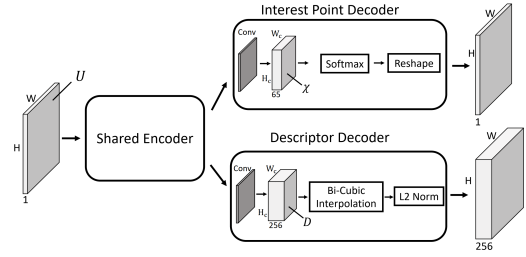


Fig. 1. SuperPoint network architecture consists of a shared encoder and two decoders.

to focus on corner points in the environment, i.e. on pixels with significant changes in pixel intensity. We use the pre-trained SuperPoint network in [12] as the teacher model in the knowledge distillation framework to train our UFEN network.

The descriptors given by SuperPoint are not binary, which makes SuperPoint unsuitable for direct integration into ORB-SLAM3 and therefore requires additional binarization operations.

C. Underwater Image Formation

In water, light reflected from the surface of underwater objects is attenuated through absorption, causing image degradation. In addition, the tiny particles in water can lead to backscattering and veiling appearance in images. An underwater image physical model is proposed in [18] to simulate these effects:

$$I_c = J_c \cdot e^{-\beta_c z} + B_c^\infty (1 - e^{-\beta_c z}). \quad (1)$$

The subscript c represents the three color channels R , G and B . z is the distance from the camera. I_c and J_c denote the pixel values of the captured underwater image and the clear in-air image, respectively. B_c^∞ is a coefficient for the background light, describing the ambient brightness in water. β_c is defined as the beam attenuation coefficient, which depends on the wavelength λ of the color channel. The parameter β_c in the formation model can be estimated through adaptive learning approaches, such as WaterGAN [19], or by referring to the Jerlov water type [20]. Jerlov water types classify different underwater scenarios based on the spectral transmittance of downward irradiance. In [18], the authors point out that the spectral transmittance in the Jerlov water type is equivalent to the *diffuse attenuation coefficient of the downwelling light* denoted by $K_d(\lambda)$.

The authors in [21] show the following relationship between $K_d(\lambda)$ and $B^\infty(\lambda)$, that is:

$$B^\infty(\lambda) = \frac{b(\lambda)E(d, \lambda)}{\beta(\lambda)}, \quad (2)$$

with

$$E(d, \lambda) = E(0, \lambda)e^{-K_d(\lambda)d}, \quad (3)$$

where $b(\lambda)$ is the beam scattering coefficient. And $E(0, \lambda)$ is the energy of light before the transmission near the water's

¹The code of UFEN-SLAM will be made public shortly.

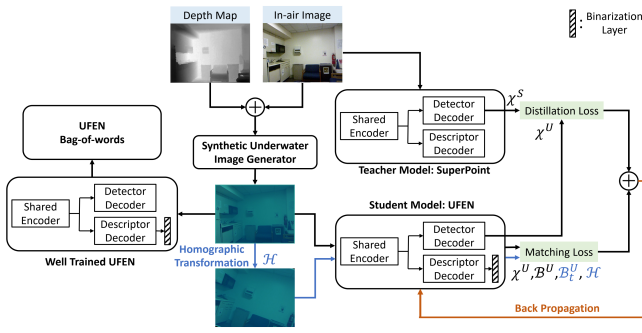


Fig. 2. Overall training framework of the UFEN. It needs the paired in-air and depth image as the training inputs. The final well-trained UFEN generates the UFEN Bag-of-words vocabulary based on the synthetic underwater images.

surface. $E(d, \lambda)$ is the energy after the transmission of a water depth d . Henceforward, $\beta(\lambda)$ describes the attenuation of the underwater beam as it transmits horizontally. $K_d(\lambda)$ governs the energy loss of the light beam propagating vertically downward in water, determining the ambient light. These two are different coefficients, but both affect underwater imaging. The values of these two coefficients for different underwater environments can be found in [20] and [22].

III. UNDERWATER FEATURE EXTRACTION NETWORK

We propose the Underwater Feature Extraction Network (UFEN) to reduce the effect of backscattering and attenuation in extracting and matching features on underwater images. Later in Section IV, we integrate the UFEN model into ORB-SLAM3. Fig. 2 shows the overall offline training framework. The first component is the synthetic underwater image generator based on a physical model using depth and in-air images (see Section III-A). To successfully integrate the trained UFEN model into the ORB-SLAM3 framework, we introduce an additional binary layer to the UFEN model. In the training of UFEN, we used the joint training of cross-modal knowledge distillation and correspondence matching to guarantee the feature extraction capability and the binary descriptor matching (see Section III-B).

A. Synthetic Underwater Image Generator

We propose to combine the physical models in (1), (2) and (3) to set a connection of the background light with the water characteristics. Then we have,

$$I_c = J_c e^{-\beta_c z} + \frac{b_c E(0, \lambda_c) e^{-K_d(\lambda_c) d}}{\beta_c} (1 - e^{-\beta_c z}) + W, \quad (4)$$

where I_c is the pixel value of the generated synthetic underwater image, determined by the wavelength λ_c of the color channel c . W is a random Gaussian noise over the image plane that simulates the flow of particles in water. The background light B_c^∞ in (1) is estimated via (2) and (3) with the coefficients $K_d(\lambda_c)$, β_c and b_c . We use the parameters suggested in [20] and [22] to generate synthetic underwater images close to the natural underwater scenarios. The value ranges of the parameters β_λ and K_d of the five open ocean

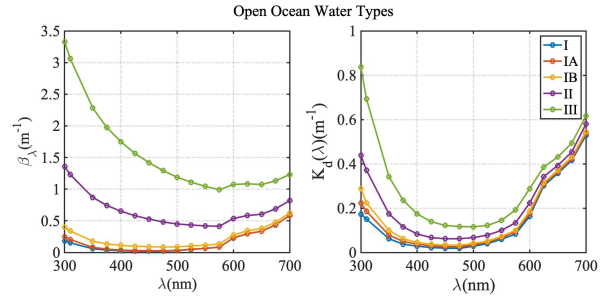


Fig. 3. The parameter values β_λ and $K_d(\lambda)$ in Open Ocean Water Types from Type I to Type III.

water types are shown in Fig. 3. We utilized water types I and III as the lower and upper bounds, and take random samples of β_λ and K_d from these intervals and apply them in (4) to generate synthetic images, to encompass the various underwater environments.

B. Underwater Feature Extraction Network (UFEN)

1) *Network Structure*: UFEN is based on the same structure as SuperPoint, with an additional binarization layer at the end of the descriptor decoder. The network input is the grayscale image $U \in \mathbb{R}^{H \times W \times 1}$. The heads of the detector and descriptor decoders compute $\mathcal{X} \in \mathbb{R}^{H_c \times W_c \times 65}$ and $\mathcal{D} \in \mathbb{R}^{H_c \times W_c \times 256}$, respectively, and finally output the tensors with the same size as the input gray image. The binarized descriptor tensor is denoted by $\mathcal{B} \in \{0, 1\}^{H \times W \times 256}$.

2) *Binary Activation Layer*: An additional binary activation layer is embedded at the end of the descriptor decoder of UFEN to fit the binary descriptor format in ORB-SLAM3. Inspired by GCNv2 [23], we use the method described in [24] to propagate the gradients through the straight-through estimator to avoid backpropagating the constant zero gradients. The binary activation function $\mathbf{b}(x)$ and the straight-through estimator g_b of the gradient are as follows:

$$\text{Forward: } \mathbf{b}(x) = \text{Sign}(x) = \begin{cases} +1 & \text{if } x \geq 0, \\ -1 & \text{Otherwise,} \end{cases} \quad (5)$$

$$\text{Backward: } g_b = 1_{|x| \leq 1}.$$

We apply the activation function in (5) to every element and get \mathcal{B} . Since the descriptor vector is L2-normalized before the binarization, it is in the range $(-1, +1)$ throughout.

3) *Loss Functions*: The loss functions aim to perform cross-modal knowledge distillation from SuperPoint to UFEN, in order to enable UFEN to achieve comparable performance on underwater images as SuperPoint does on in-air images, minimising the impact of low visibility and turbidity in water on feature detection.

The loss function used in this paper consists of two parts L_p and L_d . The superscript S and U are used to denote the outputs given by SuperPoint and UFEN, respectively. The first component L_p is for the knowledge distillation between the computed tensor of the detector decoder of SuperPoint \mathcal{X}^S and that of UFEN \mathcal{X}^U . The knowledge of binary descriptors cannot be directly distilled from SuperPoint. Therefore

we adopt the homographic transformation \mathcal{H} to simulate the change of the viewpoint and use L_d to guarantee the binary descriptor matching performance through the matching loss of corresponding feature points between images. \mathcal{B}^U and \mathcal{B}_t^U are the outputs of the UFEN descriptor decoder on the original image and the transformed images. We use the coefficient α to adjust the training weights of the two components:

$$L(\mathcal{X}^S, \mathcal{X}^U, \mathcal{B}^U, \mathcal{B}_t^U, \mathcal{H}) = L_p(\mathcal{X}^S, \mathcal{X}^U) + \alpha L_d(\mathcal{X}^U, \mathcal{B}^U, \mathcal{B}_t^U, \mathcal{H}). \quad (6)$$

The loss component L_p tends to make SuperPoint and UFEN’s detector decoder output the close probability maps and select similar feature points. To achieve this goal, we propose to use the Kullback–Leibler (KL) divergence to distil a “soft target” and adopt the Probabilistic Knowledge Transfer (PKT) [25] to distil the global feature distribution. In particular, the loss function L_p is set to be:

$$L_p(\mathcal{X}^S, \mathcal{X}^U) = L_{KL}(\mathcal{X}^S, \mathcal{X}^U) + \beta L_{PKT}(\mathcal{X}^S, \mathcal{X}^U), \quad (7)$$

with

$$L_{KL}(\mathcal{X}^S, \mathcal{X}^U) = \frac{1}{H_c W_c} \sum_{\substack{h=1 \\ w=1}}^{H_c, W_c} D_{KL}(\sigma(x_{hw}^S), \sigma(x_{hw}^U)), \quad (8)$$

in which $x_{hw}^S \in \mathbb{R}^{65}$ and $x_{hw}^U \in \mathbb{R}^{65}$ are the vectors in \mathcal{X}^S and \mathcal{X}^U , and $\sigma(\cdot)$ is the *Softmax* function that converts a real number vector to a probability distribution. The function D_{KL} is defined as,

$$D_{KL}(s, u) = \sum_{k=1}^{65} s_k \cdot (\log(s_k) - \log(u_k)), \quad (9)$$

with $s = [s_1, s_2, \dots, s_{65}]$ and $u = [u_1, u_2, \dots, u_{65}]$ ($s \geq 0$, $u \geq 0$, $\sum s_k = 1$, $\sum u_k = 1$). The KL loss quantifies the difference between two probability maps element-by-element but lacks the correlation information between the elements. Significantly, the image blur on underwater images distributes unevenly, where near places are clear and more distant places are blurred. Therefore, we use PKT loss [25] to help UFEN to have a high-level view of the feature point detection knowledge on an underwater image and learn the global information. The PKT loss distils the probabilistic knowledge from a teacher model to a student model by minimizing the divergence of the probability distribution given by the output of the two models. It uses the cosine distances between the feature vectors to reflect the distribution pattern. The PKT loss is defined as follows:

$$L_{PKT}(\mathcal{X}^S, \mathcal{X}^U) = \sum_{i=1}^{H_c W_c} \sum_{j=1, i \neq j}^{H_c W_c} s_{j|i} \log\left(\frac{s_{j|i}}{u_{j|i}}\right), \quad (10)$$

where $s_{j|i}$ and $u_{j|i}$ are the conditional probability distribution of \mathcal{X}^S and \mathcal{X}^U . It is defined as:

$$p_{i|j} = \frac{K(x_i, x_j)}{\sum_{k=1, k \neq j}^{H_c W_c} K(x_k, x_j)} \in [0, 1], \quad (11)$$

where x is the L2-normalized feature vector $\in \mathbb{R}^{65}$ extracted from \mathcal{X}^S or \mathcal{X}^U after *Softmax* and $K(x_i, x_j)$ is the similarity metric and calculated by the cosine distance between vectors,

$$K(a, b) = \frac{1}{2} \left(\frac{a^T b}{\|a\|_2 \|b\|_2} + 1 \right) \in [0, 1]. \quad (12)$$

We use the matching loss L_d to optimize the matching performance of the binary descriptors. The original image and the transformed image are the inputs of the network UFEN as shown in Fig. 2. We first find the feature points on the original image using \mathcal{X}^U , use \mathcal{H} to project these points to the transformed image, and then find all matching points between the two images. We notate the matching points on the original image and transformed image as x_i and x_i^t for $i = 1 \dots N$, respectively. We denote the descriptors corresponding to x_i and x_i^t as d_i and d_i^t , respectively. N is the number of matching points. The superscript t is used for the points and descriptors on the transformed image.

We further define the non-matching points of each x_i . Given a x_i on the original image, a point x_j^t on the transformed image is a non-matching point of x_i if $j \neq i$ and $|x_j^t - x_i^t| > T$, where T is a pre-specified constant threshold. We use $x_{i,j}^t$ to denote the non-matching points of x_i . Correspondingly, the descriptor of $x_{i,j}^t$ is denoted by $d_{i,j}^t$. Note that all d_i , d_i^t , $d_{i,j}^t$ are binary vectors with dimension equal to 1×256 extracted from \mathcal{B}^U or \mathcal{B}_t^U .

The loss L_d is similar to the loss function proposed in [26] to minimise the descriptor distance between d_i and d_i^t for all matching points, i.e., $i = 1 \dots N$ and maximise the descriptor distance between d_i and the descriptors of the non-matching points $d_{i,j}^t$. L_d is defined as:

$$L_d(\mathcal{X}^U, \mathcal{B}^U, \mathcal{B}_t^U, \mathcal{H}) = \frac{1}{N} \sum_{i=1}^N (p_i^2 + n_i^2), \quad (13)$$

where

$$\begin{aligned} p_i &= \max(0, \text{dist}(d_i, d_i^t) - P), \\ n_i &= \max(0, Q - \min(\text{dist}_n(d_i), \text{dist}_n(d_{i,j}^t))), \end{aligned} \quad (14)$$

The constants P and Q are two distance margins to the acceptable descriptor distance of matching points and non-matching points. The function $\text{dist}(\cdot)$ is the hamming distance - the number of different symbols between two binary descriptors. $\text{dist}_n(d_i)$ finds the minimal descriptor distance between d_i to the descriptors of the non-matching point set $d_{i,j}^t$,

$$\text{dist}_n(d_i) = \min(\text{dist}(d_i, d_{i,j}^t)). \quad (15)$$

IV. UFEN-SLAM

The framework of UFEN-SLAM is shown in Fig. 4. UFEN-SLAM is based on the monocular ORB-SLAM3 framework, where the trained UFEN replaces the ORB feature in ORB-SLAM3 (green blocks). UFEN has two decoders. The detector decoder extracts feature points while the other decoder assigns a 256-dimensional binary descriptor to each feature point. After detecting feature points and finding

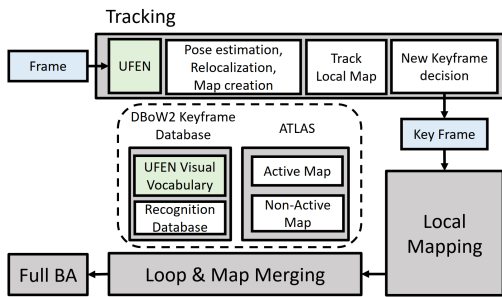


Fig. 4. UFEN-SLAM system framework

the correspondences, we follow the rest steps in ORB-SLAM3. Referring to the integration method in GCNv2 [23], we also use nearest-neighbour search matching in the reference keyframe tracking, replacing the bag-of-words matching in ORB-SLAM3, to show the actual feature extraction and matching performance of UFEN. Finally, UFEN regenerates the bag-of-words vocabulary on synthetic underwater images for loop closure detection.

V. EASI UNDERWATER DATASET

To investigate the influence of water turbidity on the performance of underwater VSLAM, we collected EASI underwater image sequences with groundtruth measured by the Vicon tracking system in the Extreme Air-Sea Interaction (EASI) Facility in Michell Hydrodynamics Laboratory at the University of Melbourne.

Experimental Setup: The EASI facility is a water tank with 1.8m width and 1.1m water depth. We installed four Vicon camera tracking systems (Vicon Vero, 2.2MP) above the tank, giving it an overlapping field-of-view area of nearly 3m length. Some green artificial plants and marine life images were fixed on a solid base plate at the bottom of the tank. We designed a rigid frame attached with a GoPro camera on one side and spatially distributed tracking markers on the other side. During the acquisition process, the camera moves in water, and the tracking markers stay above the water and are tracked by the Vicon system.

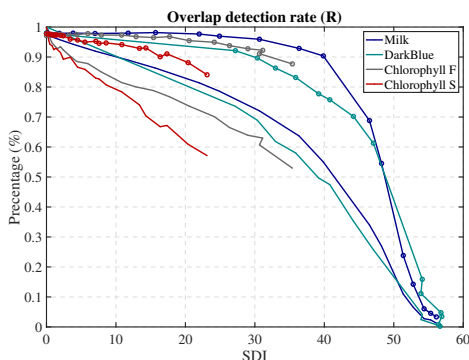


Fig. 5. Feature detection evaluation on TURBID dataset. The lines with circle markers are the results of UFEN, others are the results of SuperPoint.



Fig. 6. Image samples in EASI dataset of three turbidity levels. From left to right: Clear, Medium, Heavy.

Data Collection: The data sequences were initially collected with a GoPro Hero 10 with a sampling rate of 30 Hz and 4K resolution and downsampled to 640×480 . GoPro HyperSmooth mode was disabled when acquiring the raw videos. The Vicon system measures the ground truth with a 100 Hz acquisition rate. The dataset consists of ten sequences with three different turbidity levels ranging from 0 to 5 NTU (Nephelometric Turbidity Units). Image samples are shown in Fig. 6. We poured different amounts of talcum powder into water to control the turbidity, quantified using a YSI ProDSS turbidity meter.

VI. EXPERIMENTAL RESULTS

In this section, we present our experimental results from two perspectives to demonstrate the effectiveness of our approach. These are the evaluation of the feature points detection using the TURBID Dataset [27], and an experimental comparison based on the EASI dataset introduced in section V.

A. Training Datasets

The training data of UFEN is based on NYU depth V2 dataset [28] and DIML/CVL RGB-D indoor Dataset [29]. The depth maps in DIML/CVL indoor dataset are filled using a colour-guided depth upsampling algorithm proposed in [30]. All the images are forcibly downsampled to 640×480 .

We use the physical model in (4) when generating synthetic underwater images. For avoiding the dissipation of all image information, the maximum image depth is set to be 3m. Meanwhile the depth of the water d is assumed to be 5m and the water surface ambient light $E(0, \lambda)$ is assumed to be perfectly white. The wavelength of the colour channels B, G, R is chosen to be 450nm, 525nm and 700nm respectively.

B. Feature Detection Evaluation on TURBID dataset

This evaluation is based on an underwater colour restoration dataset - the TURBID Dataset [27], which mixes different amounts of milk, blue ink and chlorophyll to adjust the turbidity level. We exemplify the model performance of UFEN and SuperPoint by comparing their overlap detection rate of the reference feature points at different turbidity levels. The detection threshold for both models has been set to 0.01. We define the reference feature points as those detected by SuperPoint on the clear image. The overlap detection is the detected point on the turbid images, which is in a 3 by 3 pixel grid centred on the reference point. We use overlap detection rate R as the evaluation metric. That is, $R = \frac{P_c}{P_R}$. P_R is the number of the reference feature

TABLE I
ATE(M) OF VSLAMS ON EASI DATASET

Data Seq.	Data Length	LED Light	TUR (NTU)	ORB SLAM3	SuperPoint SLAM	UFEN (w/o CMD) SLAM	UFEN SLAM
01	1' 14"	On	Clear	0.0198	0.0202	0.0191	0.0196
02	0' 59"	On	(0-2)	0.0181	0.0195	0.0198	0.0205
03	1' 18"	On	Medium (2-3.5)	0.0197	0.0171	0.0174	0.0168
04	1' 32"	On		0.0178	0.0207	0.0177	0.0152
05	1' 24"	Off		F	F	0.0236	0.0203
06	1' 14"	On	Heavy (3.5-5)	0.0192	0.0226	0.0198	0.0173
07	1' 22"	On		0.0241	0.0255	0.0235	0.0234
08	1' 11"	On		0.0187	0.0185	0.0173	0.0166
09	1' 20"	On		F	F	F	0.0152
10	1' 27"	On		F	F	F	0.0193

points, and P_c is the number of the overlap detection. We adopt SDI (Structural Degradation Index) [31] to quantify the turbidity level of the images. The maximum SDI of Milk and DeepBlue is up to nearly 60, until all pattern information disappears. The SDI of chlorophyll datasets provided by TURBID is from 0 to 40. Fig. 5 visualises the varying detection performance across the different turbidity levels.

Our model UFEN has a larger R than the teacher model SuperPoint in all cases. In particular, When the SDI is less than 40, the decrease of R of UFEN is much smoother than SuperPoint. When the image is too fuzzy ($SDI \geq 50$), such as at the end of the "Milk" and "DarkBlue" data, both UFEN and SuperPoint cannot detect enough feature points and R approaches zero. The result shows that the performance of the feature detection of UFEN is less affected by the turbidity compared with SuperPoint.

C. VSLAM Evaluation on EASI Dataset

In this section, we test the proposed UFEN-SLAM using the EASI dataset, and compare it with ORBSLAM3 and SuperPoint-SLAM. In SuperPoint-SLAM, SuperPoint with an additional binarization layer is retrained using the loss function in (6), but without synthetic images.

To solve the issue of the scale uncertainty in the monocular VSLAM, we use Umeyama's method [32] to align the estimated trajectories and find the similarity transformation $S' = \{s', R', t'\}$, where R' , s' and t' are the rotational matrix, estimated scale, and translation matrix, respectively. Then the quality of the estimated trajectory is evaluated by the Absolute Trajectory Error (ATE) [33].

$$ATE_{pos} = \left(\frac{1}{N} \sum_{i=0}^{N-1} \|p_i - (s'R'\hat{p}_i + t')\|^2 \right)^{\frac{1}{2}} \quad (16)$$

where $\{p_i\}_{i=0}^{N-1}$ is the ground-truth position sequences and $\{\hat{p}_i\}_{i=0}^{N-1}$ is the estimated positions given by VSLAM algorithms. GoPro and Vicon are two separate systems, so to solve the time synchronization issue, we tune the time interval distance between two systems to find the minimal ATE.

We manually tune the thresholds in the ORB feature, to achieve the best performance of ORB-SLAM3 in turbid water. The experimental results are shown in Table

I. It shows that UFEN-SLAM performs more robust than ORB-SLAM3 and SuperPoint-SLAM by achieving lower ATEs and more successful tracking on the data sequences at the Heavy and Medium turbidity levels. ORB-SLAM3 fails to initialize on the sequence Heavy-10 and Medium-05, and fails to finish the tracking of Heavy-09. We note that Medium-05 is recorded with a darker light condition, which makes the tracking more challenging. When the water body is clear, ORB-SLAM3 and SuperPoint-SLAM show comparable performance with UFEN-SLAM. Meanwhile, to assess the efficacy of knowledge distillation, we retrained the student model without distilling the cross-modal knowledge of SuperPoint on the in-air images. Instead, we directly fed the synthetic images to SuperPoint and aimed to tune the matching performance of UFEN on the synthetic underwater images without losing performance in detection. Our results indicated that UFEN without Cross-Modal Distillation (w/o CMD) resulted in moderate improvements compared to SuperPoint-SLAM on Medium and Heavy levels. It has lower ATEs on Heavy sequences than SuperPoint-SLAM, and unlike SuperPoint-SLAM and ORB-SLAM3, UFEN (w/o CMD) SLAM can track the Medium-05 sequence successfully. However, UFEN (w/o CMD) SLAM again fails to track Heavy-09 and Heavy-10 and has higher ATEs than UFEN-SLAM on both Heavy and Medium-level sequences. These results demonstrate that UFEN without Cross-Modal Distillation can still slightly improve underwater tracking performance compared to the teacher model. To summarise, Cross-Modal Distillation can make the model more robust in a turbid underwater environment. That proves the effectiveness of Cross-Modal Distillation in our training.

VII. CONCLUSION

This paper proposes a monocular UFEN-VSLAM for turbid underwater scenes. Moreover, we collect the EASI underwater dataset with multiple turbidity levels. Our approach performs effectively on TURBID and EASI datasets.

ACKNOWLEDGMENTS

Authors thank Mr. Max Rounds for the technical support in the experiments. Authors gratefully acknowledge funding support from Australian government grants AUS-MURIB000001, DE220101527 and DE210101624.

REFERENCES

- [1] J. González-García, A. Gómez-Espinoza, E. Cuan-Urquizo, L. G. García-Valdovinos, T. Salgado-Jiménez, and J. A. Escobedo Cabello, "Autonomous underwater vehicles: Localization, navigation, and communication for collaborative missions," *Applied sciences*, vol. 10, no. 4, p. 1256, 2020.
- [2] A. Sahoo, S. K. Dwivedy, and P. Robi, "Advancements in the field of autonomous underwater vehicle," *Ocean Engineering*, vol. 181, pp. 145–160, 2019.
- [3] J. Cheng, L. Zhang, Q. Chen, X. Hu, and J. Cai, "A review of visual slam methods for autonomous driving vehicles," *Engineering Applications of Artificial Intelligence*, vol. 114, p. 104992, 2022.
- [4] C. Duan, S. Junginger, J. Huang, K. Jin, and K. Thurow, "Deep learning for visual slam in transportation robotics: a review," *Transportation Safety and Environment*, vol. 1, no. 3, pp. 177–184, 2019.
- [5] B. Joshi, S. Rahman, M. Kalaitzakis, B. Cain, J. Johnson, M. Xanthidis, N. Karapetyan, A. Hernandez, A. Q. Li, N. Vitzilaios *et al.*, "Experimental comparison of open source visual-inertial-based state estimation algorithms in the underwater domain," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 7227–7233.
- [6] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [7] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.
- [8] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.
- [9] R. Garcia and N. Gracías, "Detection of interest points in turbid underwater images," in *OCEANS 2011 IEEE-Spain*. IEEE, 2011, pp. 1–9.
- [10] Z. Luo, L. Zhou, X. Bai, H. Chen, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan, "Aslfeat: Learning local features of accurate shape and localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6589–6598.
- [11] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger, "R2d2: repeatable and reliable detector and descriptor," *arXiv preprint arXiv:1906.06195*, 2019.
- [12] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.
- [13] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5173–5182.
- [14] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, pp. 1150–1157.
- [15] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.
- [16] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2827–2836.
- [17] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.
- [18] D. Akkaynak and T. Treibitz, "A revised underwater image formation model," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6723–6732.
- [19] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robotics and Automation letters*, vol. 3, no. 1, pp. 387–394, 2017.
- [20] N. Jerlov, "Optical oceanography," 1968.
- [21] D. Akkaynak and T. Treibitz, "Sea-thru: A method for removing water from underwater images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1682–1691.
- [22] M. G. Solonenko and C. D. Mobley, "Inherent optical properties of jerlov water types," *Applied optics*, vol. 54, no. 17, pp. 5392–5401, 2015.
- [23] J. Tang, L. Ericson, J. Folkesson, and P. Jensfelt, "Gcnv2: Efficient correspondence prediction for real-time slam," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3505–3512, 2019.
- [24] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1," *arXiv preprint arXiv:1602.02830*, 2016.
- [25] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 268–284.
- [26] R. Pautrat, V. Larsson, M. R. Oswald, and M. Pollefeys, "Online invariance selection for local feature descriptors," in *European Conference on Computer Vision*. Springer, 2020, pp. 707–724.
- [27] A. Duarte, F. Codevilla, J. D. O. Gaya, and S. S. Botelho, "A dataset to evaluate underwater image restoration methods," in *OCEANS 2016-Shanghai*. IEEE, 2016, pp. 1–6.
- [28] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgb-d images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.
- [29] J. Cho, D. Min, Y. Kim, and K. Sohn, "Dm1/cv1 rgb-d dataset: 2m rgb-d images of natural indoor and outdoor scenes," *arXiv preprint arXiv:2110.11590*, 2021.
- [30] Y. Kim, B. Ham, C. Oh, and K. Sohn, "Structure selective depth superresolution for rgb-d cameras," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5227–5238, 2016.
- [31] R. Garcia and N. Gracías, "Detection of interest points in turbid underwater images," in *OCEANS 2011 IEEE-Spain*. IEEE, 2011, pp. 1–9.
- [32] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 04, pp. 376–380, 1991.
- [33] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 7244–7251.