Unsupervised Learning-Based Optical–Acoustic Fusion Interest Point Detector for AUV Near-Field Exploration of Hydrothermal Areas

Yihui Liu 1, Yufei Xu 1,*, Ziyang Zhang 1, Lei Wan 1, Jiyong Li 2 and Yinghao Zhang 3

1 Science and Technology on Underwater Vehicle Laboratory, Harbin Engineering University, Harbin 150001, China; liuyihuiheu@hrbeu.edu.cn (Y.L.); zhangziyang@hrbeu.edu.cn (Z.Z.); wanlei@hrbeu.edu.cn (L.W.)
2 Tianjin Navigation Instruments Research Institute, Tianjin 300131, China; lijiyong2022@163.com
3 Wuhan Second Ship Design and Research Institute, Wuhan 430063, China; zhangyinghao555@126.com
* Correspondence: xuyufei@hrbeu.edu.cn

Abstract: The simultaneous localization and mapping (SLAM) technique provides long-term near-seafloor navigation for autonomous underwater vehicles (AUVs). However, the stability of the interest point detector (IPD) remains challenging in the seafloor environment. This paper proposes an optical–acoustic fusion interest point detector (OAF-IPD) using a monocular camera and forward-looking sonar. Unlike the artificial feature detectors most underwater IPDs adopt, a deep neural network model based on unsupervised interest point detector (UnsuperPoint) was built to reach stronger environmental adaption. First, a feature fusion module based on feature pyramid networks (FPNs) and a depth module were integrated into the system to ensure a uniform distribution of interest points in depth for improved localization accuracy. Second, a self-supervised training procedure was developed to adapt the OAF-IPD for unsupervised training. This procedure included an auto-encoder framework for the sonar data encoder, a ground truth depth generation framework for the depth module, and optical–acoustic mutual supervision for the fuse module training. Third, a non-rigid feature filter was implemented in the camera data encoder to mitigate the interference from non-rigid structural objects, such as smoke emitted from active vents in hydrothermal areas. Evaluations were conducted using open-source datasets as well as a dataset captured by the research team of this paper from pool experiments to prove the robustness and accuracy of the newly proposed method.

Keywords: underwater localization; sensors fusion; unsupervised learning; SLAM

1. Introduction

Near-field exploration of hydrothermal fields provides intricate visual data that are essential for marine geological and biological research [1,2]. Nevertheless, achieving precise localization of autonomous underwater vehicles (AUVs) in unstructured and constrained environments remains challenging. Recent advancements in simultaneous localization and mapping (SLAM) technologies, driven by progress in autonomous driving and unmanned aerial vehicles and virtual or augmented reality, offer new possibilities. Bathymetric SLAM (BLSAM) technology, which uses multi-beam sonar to measure seabed topography, has been proposed to enable long-term navigation for AUVs but is unsuitable for close-range observation [3]. Visual SLAM (VLSAM) technology, on the other hand, employs cameras to collect environmental data and estimate the vehicle’s pose through inter-frame information [4]. This method simultaneously creates an environmental map and determines the vehicle’s position on it, facilitating self-navigation and localization.

However, studies on open-source algorithms reveal that in environments with poor textures or inconspicuous features, pure visual interest point detector (IPD) and VLSAM systems suffer from significant position errors [5–7]. As shown in Figure 1, common visual
challenges in underwater environments include inadequate light, low image contrast, and color attenuation, all of which adversely affect AUVs’ visual positioning [8,9]. Moreover, scale estimation with monocular cameras remains problematic, and using stereo cameras to provide scale through image matching can increase computational demands and space usage [10]. Certain methods that integrate other sensors, such as sonar or depth meters, have achieved some success in underwater visual–inertial odometry (VIO) [11–13]. Therefore, additional sonar data are essential for our IPD method to accurately estimate AUV posture changes.

The performance of a feature-based IPD largely depends on the number and geometric distribution of features obtained by the feature extraction and description algorithm. However, the IPDs most used in underwater SLAM, such as scale-invariant feature transform (SIFT), speeded-up robust features (SURFs), and oriented FAST and rotated BRIEF (ORB), are designed for air-mediated environments [14–16]. These detectors often capture fewer interest points due to the adverse underwater imaging environment [17]. On the other hand, descriptors generated by these detectors frequently fail to match correctly because of the intense variations in illumination from AUV-equipped lighting. Deep neural networks, with their superior environmental adaptation and illumination robustness, outperform traditional image processing methods in many challenges [18–20]. Consequently, a deep learning framework for near-field exploration missions in hydrothermal fields is more appropriate.

In light of the aforementioned restrictions, an optical–acoustic fusion interest point detector (OAF-IPD) system, as shown in Figure 2, comprising a monocular camera and a forward-looking sonar was proposed. A deep neural network based on an unsupervised interest point detector named UnsuperPoint forms the foundation, providing features that are more suitable for the near-seafloor environment when compared to traditional methods such as ORB and FAST [20]. To establish the relevance between sensors, a feature fuse module based on feature pyramid networks (FPNs) was introduced to the OAF-IPD [21]. For better localization accuracy, the depth module was designed to ensure a uniform distribution of interest points in depth. It is important to note that the term “depth” used in this paper does not refer to the water depth at which the vehicle is operating. Instead, it refers to the distance between each pixel in the image and its corresponding point in the real world. To realize unsupervised training, this paper designed a training strategy

Figure 1. Examples of underwater images that are not conducive to feature point selection and matching; (a) few features can be extracted, (b) varying illumination conditions, (c) non-rigid objects in yellow boxes (such as black chimneys or underwater organisms).
containing an auto-encoder framework for sonar data encoder training, a ground truth depth generation framework for depth module training, and optical–acoustic mutual supervision for fuse module training. Considering the accuracy impact caused by the features from non-rigid structural objects, a non-rigid feature removal training framework was designed.

To substantially improve the robustness and accuracy of AUV self-localization in the hydrothermal near-seafloor environment, OAF-IPD addresses the limitations of individual sensors by fusing multiple sensor data with the deep learning framework. OAF-IPD aims to realize a more accurate unsupervised learning IPD method works robustly in hydrothermal near-seafloor environment. The key contributions are as follows:

1. Fuse module incorporating the FPN into UnsuperPoint for better sensor fusion;
2. Depth module designed to ensure a uniform distribution of interest points in depth for better localization accuracy;
3. Unsupervised training strategy including an auto-encoder framework, a ground truth depth generation framework, and a mutually supervised framework designed to enable unsupervised training of the new modules;
4. Non-rigid feature filter with which the camera data encoder filters out features from non-rigid structural objects, mitigating the interference caused by the smoke emitted from active vents in hydrothermal areas;

The paper is structured as follows: Section 2 provides a brief introduction to related works; Section 3 offers the network architecture of the proposed OAF-IPD method; Section 4 details the self-supervised frameworks of OAF-IPD; Section 5 is the design of the loss function; Section 6 presents results from a diverse set of challenging underwater environments. Finally, Section 7 summarizes the conclusions and directions for future work.

Figure 2. Overview of OAF-IPD.
function, and Section 6 presents results from a diverse set of challenging underwater environments. Finally, Section 7 summarizes the conclusions and directions for future work.

2. Related Works

2.1. IPD for Underwater SLAM

As a critical component of SLAM, the interest point detector has been extensively researched, leading to the development of numerous feature point selection and description algorithms tailored for underwater visual SLAM.

The monocular SLAM method used in AUV hull inspections leverages the SIFT and Harris feature detectors for image registration to calculate relative camera poses [22]. This method was validated using 1300 images from a U.S. aircraft carrier hull, demonstrating the capability for automatic mapping and navigation of the hull’s underwater surface.

A multi-stage visual odometry system with a fault detection mechanism was developed to handle motion measurement and correct unreliable estimates in texture-poor underwater environments [23]. Using a monocular camera, the system includes a primary motion estimation based on the scene features and two backup systems. Using a monocular camera, the system includes a primary motion estimation based on scene features and two backup systems: one estimating the AUV motion via phase correlation and the other using high-order motion prediction.

2.2. Deep Learning Based IPD

Traditional feature-based IPD methods will significantly underperform in underwater imaging conditions. To improve the robustness of the feature extraction and description algorithm, this paper attempts to apply a deep learning-based feature extraction and description method to the underwater images and sonar data.

Learned invariant feature transform (LIFT) predicts both points and descriptors using three modules: a detector that creates a score map of interest points, an orientation estimator that predicts patch orientations, and a descriptor module [18]. The score map is used to crop patches of interest points. A spatial transformer network (STN) [24] rotates patches according to the estimated orientation before generating descriptors. LIFT uses image patches as input to effectively learn localized feature detection and description. The model requires an initial structure-from-motion (SfM) pipeline to generate point cloud data for supervision to guide the training of each component. The different components of LIFT utilize deep learning techniques but are not jointly trained within an end-to-end framework. As a result, the training and inference processes are conducted separately. Each module operates independently, making LIFT unsuitable for real-time application due to its slow speed.

SuperPoint is an unsupervised deep learning-based interest point detector and descriptor [19]. It addresses the challenge of creating consistent ground truth data for interest points by using a self-supervised Siamese network with a novel loss function, enabling automatic learning of interest point scores and positions. SuperPoint simplifies the training process by not requiring pseudo ground truth or structure-from-motion representations. This allows for end-to-end training in a single round, delivering real-time performance and reliable accuracy.

UnsuperPoint, an end-to-end unsupervised interest point detector and descriptor, was proposed for interest point detection and descriptor extraction [20]. It employs self-supervised learning, leveraging geometric transformations between images to eliminate the need for labeled data and relying on geometric consistency for training. The advantages of UnsuperPoint include scalability, robustness to geometric transformations, and high-quality, consistent descriptors across views, enhancing performance in SLAM, 3D reconstruction, and augmented reality.
3. Network Architecture

This section details the network architecture of the OAF-IPD. The overall structure of the proposed OAF-IPD system is illustrated in Figure 3. UnsuperPoint is considered as the basis of OAF-IPD due to its learnable features and unsupervised learning capabilities. Additionally, a sonar data processing component, a feature fusion component, and a depth prediction module were integrated into the OAF-IPD.

The OAF-IPD inputs from the monocular camera images and forward-looking multibeam sonar data. The output is divided into four parts: a score map of interest points, a point location map, a depth map, and a descriptor map.

3.1. Data Encoders

The data encoders, encompassing both camera and sonar data processing components, can autonomously learn features without manual specification and can be specifically trained using underwater images. This capability enhances their adaptability to underwater conditions compared to manually designed features such as ORB and SIFT.
3.1.1. Backbone

This paper adopts a backbone the same as UnsuperPoint to process the input data from the camera or sonar, generating an intermediate feature map for subsequent subtasks. This backbone is a fully convolutional network comprising four pairs of convolutional layers, interspersed with four max-pooling layers, each with a stride and kernel size of two. After each pooling layer is the number of channels in the subsequent double convolutional layers; specifically, the channel configuration across the ten convolutional layers is 32-32-64-64-128-128-256-256-256-256. All convolutional layers are configured with a stride of 1 and a kernel size of 3. Each convolutional layer, except for the final one in each subtask, is followed by batch normalization and a leaky ReLU activation function [25,26]. The entries are processed through a fully convolutional approach for each subtask, enabling the module to manage input images of different sizes [27].

3.1.2. Camera Data Encoder

The function of the camera data encoder is to map the input image $I_c \in \mathbb{R}^{H \times W \times 3}$ to the mid-layer tensor $F_{\text{camera}} \in \mathbb{R}^{H_c \times W_c \times T}$. Utilizing the aforementioned backbone, this tensor achieves reduced spatial dimensions but increased channels (i.e., $H_c < H$, $W_c < W$, and $T > 3$). This transformation captures high-level information while reducing dimensionality. In this paper, $f_{\text{scale}} = 4, 8, 16$ is adopted for three different scales of the mid-layer tensor $F_{\text{camera}}$, $W_c = W / f_{\text{scale}}$, and $H_c = H / f_{\text{scale}}$.

3.1.3. Sonar Data Encoder

Sonar data are recorded in polar coordinates, leading to non-uniform spatial distribution – dense point clouds near the sonar and sparse ones further away. This non-uniformity complicates the subsequent data processing. The input sonar data are defined as $I_s \in \mathbb{R}^{D \times N}$:

$$
\begin{align*}
D &= R / \Delta r \\
N &= \psi / \Delta \psi
\end{align*}
$$

where $R$ is the sonar detection distance, $\Delta r$ is radial resolution. $\psi$ is the sonar detection angle range and $\Delta \psi$ is angular resolution.

A smaller network is utilized as the sonar data encoder, which is similar to the backbone network in Section 3.1.1 but without the last two 256-kernel convolutional layers and a pooling layer. The sonar data encoder generates the mid-layer tensor $F_{\text{sonar}} \in \mathbb{R}^{H_s \times W_s \times T}$, $H_s = H / 8$, and $W_s = W / 8$.

3.2. Fuse Module

As shown in Figure 4, the fuse module integrates the camera and sonar data extracted by the encoders. Utilizing the feature pyramid network (FPN), the mid-layer tensors $F_{\text{camera}}$ with $f_{\text{scale}} = 4, 8, 16$ are fused [21]. To achieve uniform kernel numbers across features from different scales in the FPN, a $1 \times 1$ convolutional cell is employed. Additionally, a “$2 \times \text{up-cell}$” is utilized to upsample features to twice their original height and width. The three resulting scales of features are concatenated and max-pooled to form a multiple-scale feature map $F_{\text{MS}} \in \mathbb{R}^{H_s \times W_s \times T}$ with scale of $f_{\text{scale}} = 8$. Finally, the fused feature map $F_{\text{fuse}}$ is obtained by adding $\lambda_{\text{son}} \times F_{\text{sonar}}$ to $F_{\text{MS}}$, where $\lambda_{\text{son}}$ is a predefined factor to balance the contributions of $F_{\text{camera}}$ and $F_{\text{sonar}}$ in the $F_{\text{fuse}}$.

This module merges feature maps from different layers, combining contextual information to generate richer and more distinguishable feature representations. This multi-scale fusion results in feature maps with enhanced representational capabilities, thereby improving the model’s performance in subsequent modules.
pooling layers (subsampling factor of 8), the module predicts a relative position for each to constrain the predictions \( D \) respectively, outputting the depth predictions \( F \). Using obtained from the sonar data does not encompass the depth information for all points. The sparse point cloud distance information the camera to the candidate interest points. The sparse point cloud distance information

3.4. Depth Module

The depth module contains two convolutional layers with 256 and 1 channels, respectively, outputting the depth predictions \( D_{pdt} \). The final layer includes a sigmoid activation to constrain the predictions \( D_{pdt} \) within the interval \([0, 1]\).
The output of the depth module $D_{\text{map}}$ is calculated as follows:

$$D_{\text{map}} = d_{\text{max}}D_{\text{pdt}}$$  \hspace{1cm} (3)

where $d_{\text{max}}$ is a parameter that represents the maximum detection distance.

3.5. Score Module

The aim of the score module is to evaluate the candidate points of interest. Points that are repeatable and reliable receive higher scores. Our score module has the same structure as the depth module, containing 256-1 convolutional layers and sigmoid activation. The output score predictions $S_{\text{map}}$ are constrained within the interval $[0, 1]$.

To ensure a uniform depth distribution of interest points, this paper calculates $S_{\text{sort}}$ by using the $D_{\text{map}}$ obtained from the depth module and using it to influence the scoring criterion. The formula is as follows:

$$S_{\text{sort}} = S_{\text{map}} + \lambda_{\text{dep}}D_{\text{map}} - \mu_{\text{dep}}$$  \hspace{1cm} (4)

where $\mu_{\text{dep}}$ is the mean depth matrix obtained by calculating the average depth for each candidate interest point and its surrounding eight interest points. $\lambda_{\text{dep}}$ is a factor that controls the weight of how the score is affected by depth.

Finally, the top $N_{\text{sort}}$ points are selected as interest points based on the sorted $S_{\text{sort}}$ after converting it into a vector.

3.6. Descriptor Module

The interest point descriptor module generates descriptors for each interest point for subsequent matching. In underwater environments, where feature point numbers are limited and exhibit minimal differences, descriptors do not require high dimensionality for representation but do require increased distinctiveness. Compared to UnsuperPoint, our method enhances the model’s depth while reducing the output dimensionality. This adjustment significantly decreases the computational load during the subsequent matching process by generating more efficient interest point descriptors. This module consists of four convolutional layers: the first and the second have 256 channels, and the next two layers reduce the channels to the target descriptor dimension of 128. Each convolutional layer is followed by batch normalization and dropout to enhance training stability and prevent overfitting [25].

Additionally, by integrating features from both the camera and sonar data, the generated descriptors are better suited for multimodal matching tasks, enhancing the accuracy and reliability. The module uses all point positions in the $P_{\text{map}}$ to interpolate all entries in the descriptor map $F_{\text{map}}$. By regressing the point positions, interpolation of descriptors becomes differentiable and is thus utilized during training.

4. Self-Supervised Framework

The optical–acoustic model mutual supervision training is proposed to enable unsupervised training of the whole model. The training framework of the camera data encoder follows the basic architecture of UnsuperPoint, with a self-supervised training framework added to eliminate non-rigid objects. Moreover, an audio–visual mutual supervision training framework based on sensor differences is proposed to enable mutual enhancement of the two encoders. The involvement of each module in each training procedure of the OAF-IPD is shown in Figure 5. The red cells indicate the corresponding modules with updated weights during specific training procedures, the light red cells denote modules with fixed weights to support the training of others, and the grey cells represent modules that were not involved in the particular training procedure.
Additionally, during vehicle movement, the intensity of a spatial point can vary significantly. To predict interest points from two augmentations of the same input image. The homograph forms the basis of the training for the camera data processing network.

Branch structure. These augmentations and their predictions are processed in separate branches: (blue) handles the unwrapped input image, while (red) processes the wrapped images which are transformed by the homography matrix. This method forms the basis of the training for the camera data processing network.

The mutually supervised framework is designed to address the shortcomings of using sonar and camera sensors independently. Due to the sparse distribution characteristics of sonar data, standalone self-supervised training of this module often leads to model non-convergence. On the other hand, direct joint training of both the sonar and camera modules can cause the model to be overly reliant on the camera data. In real underwater operational environments, where the AUV provides the only available light source, images suffer from highly uneven illumination, leading to the clustering of high-scoring feature points. Additionally, during vehicle movement, the intensity of a spatial point can vary significantly between frames, adversely affecting the interest point generation by the camera.

4.1. Ground Truth Generation

To achieve unsupervised training, the OAF-IPD needs to use unlabeled sonar and image data and employ non-manual methods to generate the required ground truth, including point pairs similar to those needed by UnsuperPoint, and additional depth maps.

4.1.1. Homography-Based Self-Supervised Framework

UnsuperPoint employs a self-supervised training framework to learn three tasks simultaneously. As shown in Figure 6, UnsuperPoint operates within a Siamese network to predict interest points from two augmentations of the same input image. The homograph branch shares weights with the camera data encoder by utilizing a Siamese network structure. These augmentations and their predictions are processed in separate branches: (blue) handles the unwrapped input image, while (red) processes the wrapped images which are transformed by the homography matrix. This method forms the basis of the training for the camera data processing network.
To establish the point correspondences (point pairs) between the two branches $A$ and $B$, each branch $b \in \{A, B\}$ outputs four tensors, $P_b$, $D_b$, $S_b$, and $F_b$, containing point positions, depths, scores, and descriptors. A distance matrix $G$ is computed using the pairwise Euclidean distances between points from branch $A$ (transformed by homography matrix $T_b$) and points from branch $B$.

$$G = \left[ S_{ij} \right]_{M_A \times M_B} = \left[ \left\| p_i^{A\rightarrow B} - p_j^B \right\|^2 \right]_{M_A \times M_B}$$

(5)

Here, $M_A$ and $M_B$ represent the number of points in branch $A$ and branch $B$, respectively. $p_i^{A\rightarrow B}$ is the $i$-th point in branch $A$, transformed by $T_b$ to align with branch $B$. $p_j^B$ is the $j$-th point in branch $B$. $\left\| p_i^{A\rightarrow B} - p_j^B \right\|^2$ denotes the squared Euclidean distance between the transformed $i$-th point from branch $A$ and the $j$-th point from branch $B$.

A point-pair is formed if a point in branch $B$ has a nearest neighbor in branch $B$ within a predefined minimum distance $\varepsilon_c$. These correspondences redefine the output tensors to new corresponding tensors $\hat{P}_b$, $\hat{S}_b$, $\hat{D}_b$, and $\hat{F}_b$ with $K$ entries. The point-pair correspondence distance $\text{Dis}_k$ is the Euclidean distance between the corresponding points from branch $A$ and $B$

$$\text{Dis}_k = \left\| T_b \hat{p}^A_k - \hat{p}^B_k \right\|$$

(6)

where $\hat{p}^A_k$ and $\hat{p}^B_k$ are the coordinates of the predicted point pair.

4.1.2. Depth Ground-Truth Generation

Through a sonar–camera joint data collection experiment conducted in a pool environment (detailed in Section 6), a combined dataset of image data and sonar data was obtained. COLMAP 3.11 is a state-of-the-art SfM and multi-view stereo (MVS) software that reconstructs 3D models from images [28]. This section used COLMAP with a feature metric refinement method using image sequence to simulate the three-dimensional (3D) reconstruction of the hydrothermal field scene and generate the vehicle trajectories [29]. Adjustments to the scale and accuracy of the reconstruction scene and trajectories were made based on the reference markers painted at the bottom of the experimental pool. The accuracy of the reconstruction can reach the centimeter level, while the accuracy of the trajectories can reach the decimeter level. Figure 7 shows the 3D reconstruction of the hydrothermal area scene generated by COLMAP.
The pixel coordinate position of the candidate interest point $p_c = (i, j)$ can be predicted through the position module:

$$P_{map}(p_c) = (u, v)$$

(7)

Define the camera intrinsic matrix $K$:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

(8)

where $f_x$, $f_y$, $c_x$, and $c_y$ can be obtained through Zhang’s method, which is used as a camera calibration method in this paper.

The direction vector $d_{cam}$ in the camera’s 3D coordinates is then given by

$$d_{cam} = K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}$$

(9)

The corresponding direction vector $d_{rec}$ of $d_{cam}$ in the 3D reconstruction coordinate system is

$$d_{rec} = Rd_{cam} + t$$

(10)
The $3 \times 3$ rotation matrix $\mathbf{R}$ and the $3 \times 1$ translation vector $\mathbf{t}$ is also the location and pose of the camera. They can be obtained through the refined COLMAP. The vector from camera position $\mathbf{t}$ to point $\mathbf{p}_{clu}$ is defined as $\mathbf{d}_p$:

$$
\mathbf{d}_p = \mathbf{p}_{clu} - \mathbf{t}
$$

$\theta$ is the angle between vector $\mathbf{d}_{rec}$ and vector $\mathbf{d}_p$:

$$
\theta = \arccos \left( \frac{\mathbf{d}_{rec} \cdot \mathbf{d}_p}{\|\mathbf{d}_{rec}\| \|\mathbf{d}_p\|} \right)
$$

$P_{cloud} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\}$ is point cloud created by the refined COLMAP. Using the K-D tree method, $\mathbf{p}_w$ is filtered out of $P_{cloud}$ by restraining the detection distance $d_{max}$ and angular range $\epsilon_\theta$ centered around $\mathbf{d}_{rec}$:

$$
\mathbf{p}_w \in \{\mathbf{p}_{cloud} \mid (\theta \leq \epsilon_\theta) \land (\|\mathbf{d}_p\| \leq d_{max}) \}
$$

For each point within this range, the distance to the camera position is calculated. The minimum distance $\|\mathbf{d}_p\|$ is taken as the ground truth depth map $D_{gt}(i, j)$ of the candidate interest point $\mathbf{p}_c$.

4.2. Sonar Encoder Pertain Based on Auto-Encoder

Before formal training, an autoencoder structure shown in Figure 8 was utilized to perform unsupervised pre-training on the encoder modules. The sonar data encoder, which uses a single-channel input backbone without a publicly available pre-trained model, was trained on the sonar dataset collected from the pool experiments.

![Figure 8](image.png)

Figure 8. Auto-encoder-based training framework.

4.3. Non-Rigid Feature Removal Training Framework

The camera data processing component of the OAF-IPD added an unsupervised training structure to handle non-rigid objects, which in hydrothermal areas primarily consisted of the smoke emitted from vents. The model was trained using image sequences of active vents captured by a stationary camera. $I_m$ and $I_n$ are images from the sequence...
\{I_1, I_2, \ldots, I_e\}$ with $\varepsilon$ images. $\varepsilon$ is a predefined step size which represents the number of frames between $I_m$ and $I_n$:

$$n = m + \varepsilon$$

The purpose of this framework is to reduce the scores of interest points with significant positional changes when the camera and rigid structural objects are stationary, relative to each other. In this context, $P_m, D_m, S_m,$ and $F_m$ are the information of interest points generated from $I_m$, and $P_n, D_n, S_n,$ and $F_n$ are the information of interest points generated from $I_m$. Feature descriptors, denoted as $f_i$ from $F_n$ and $f_j$ from $F_n$, are used to identify corresponding points by calculating the distance $D_f(i,j)$

$$D_f(i,j) = \|f_i - f_j\|_2$$

$\varepsilon_f$ is the predefined minimum distance for correspondence determination. Point pairs are set to the corresponding point pairs when $D_f(i,j) < \varepsilon_f$.

The model aims to minimize the distance between the positions of the $k$-th corresponding point pair, represented as

$$D_p(k) = \|P_m(k) - P_n(k)\|_2$$

After this specialized optimization, the score module will assign low scores to points from non-rigid objects.

### 4.4. Mutually Supervised Framework Based on Sensor Differences

The mutually supervised framework involves a mutual training process between the camera data encoder and the sonar data encoder. With the camera initially supervising the training of the sonar and the sonar subsequently supervising the camera, the advantages of both sensors are fully leveraged.

#### 4.4.1. Camera-Supervised Framework

To address the issue of sparse spatial distribution in the sonar data, this paper reinforces the sonar encoder using the previously trained camera data encoder. By inputting a well-illuminated image with additional lighting, the camera model can reliably generate interest points. These interest points serve as ground truth to train the sonar data encoder with updates applied solely to the sonar encoder’s weights while keeping other modules fixed.

#### 4.4.2. Sonar-Supervised Framework

To mitigate the impact of poor underwater lighting conditions on the camera, interest points generated from sonar data supervise the camera data encoder’s training, enhancing its adaptability to adverse lighting conditions.

### 5. Loss Functions

By adding a depth prediction loss $L_{depth}$ and a non-rigid feature loss $L_{rig}$, the OAF-IPD improves upon the four loss functions used in the UnsuperPoint for scores, positions, and descriptor training. The total loss $L_{total}$ is expressed as

$$L_{total} = \alpha_s L_s + \alpha_p L_p + \alpha_{depth} L_{depth} + \alpha_{desc} L_{desc} + \alpha_{rpt} L_{rpt} + \alpha_{rig} L_{rig}$$

Each term is weighted by a factor $\alpha$. $L_s$ is the score module loss to ensure similar scores for corresponding points $k$ in sets $A$ and $B$.

$$L_s = \sum_{k=1}^{K} \left( \hat{s}^A_k - \hat{s}^B_k \right)^2$$
where $\hat{s}_A^k$ and $\hat{s}_B^k$ are the scores of the corresponding points $k$ in sets $A$ and $B$, respectively. $K$ is the total number of corresponding points.

The objective of the second loss $L_p$ term is to ensure that the predicted positions of the interest points represent the same point in the image pair:

$$L_p = \sum_{k=1}^{K} Dis_k$$  \hspace{1cm} (19)

where $Dis_k$ is defined as the distance of the corresponding point pair in Equation (6).

$L_{rpt}$ represents the repeatability of points.

$$L_{rpt} = \sum_{k=1}^{K} (\hat{s}_A^k + \hat{s}_B^k) \cdot (Dis_k - \mu_{dis})$$  \hspace{1cm} (20)

where $\mu_{dis}$ denotes the average distance between all point pairs.

$d_{gl}^k$ is the average ground truth depth of the point pairs $k$:

$$d_{gl}^k = \frac{D_{gt}(\hat{p}_A^k) + D_{gt}(\hat{p}_B^k)}{2}$$  \hspace{1cm} (21)

where $D_{gt}(i,j)$ is the ground truth depth map generated in Section 4.1.2. $\hat{p}_A^k$ and $\hat{p}_B^k$ are the coordinate values of the predicted point pair.

$L_{depth}$ represents the loss for the point pair depth disparity:

$$L_{depth} = \sum_{k=1}^{K} \frac{(\hat{d}_A^k - d_{gl}^k)^2 + (\hat{d}_B^k - d_{gl}^k)^2}{2}$$  \hspace{1cm} (22)

$\hat{d}_A^k$ and $\hat{d}_B^k$ are the predicted depths of corresponding points $k$ in sets $A$ and $B$, respectively.

$L_{desc}$ is expressed as Equation (23) to ensure that the descriptors for corresponding points are similar, while the descriptors for non-corresponding points are dissimilar. The descriptor loss is determined using a hinge loss with a positive and a negative margin, as described in SuperPoint and UnsuperPoint [19,20].

$e_{ij}$ is an indicator function that represents the correspondence of points, where points more than 8 pixels apart increase the loss.

$$L_{desc} = \frac{M_A \cdot M_B}{\sum_{i=1}^{M_A} \sum_{j=1}^{M_B} [\lambda_{desc} \cdot e_{ij} \cdot \max(0, m_p - f_i^A T f_i^B) + (1 - e_{ij}) \cdot \max(0, f_i^A T f_i^B - m_n)]}$$  \hspace{1cm} (23)

$$e_{ij} = \begin{cases} 0, & g_{ij} \leq 8 \\ 1, & g_{ij} > 8 \end{cases}$$  \hspace{1cm} (24)

where $M_A$ and $M_B$ represent the number of points in sets $A$ and $B$, $m_p$ and $m_n$ are the margins for positive and negative pairs in the hinge loss. $f_i^A$ and $f_i^B$ are the feature descriptors for points $i$ in set $A$ and $j$ in set $B$. A weight term $\lambda_{desc}$ is added to balance the few corresponding points compared to the non-corresponding points.

$L_{rig}$ is only used in specialized training for non-rigid interest point removal:

$$L_{rig} = \frac{1}{q} \cdot \sum_{k=1}^{K} D_p(k)$$  \hspace{1cm} (25)
6. Experiments and Results

6.1. Data Collection

To train the proposed IPD method, it was essential to simultaneously acquire data from both the camera and sonar. Using the models depicted in Figure 9b,c,d, a simulation environment of hydrothermal fields was constructed in an experimental pool shown in Figure 9e. The sonar-camera combined dataset was collected with the Blue-ROV, as shown in Figure 9a. Image sequences and sonar data were simultaneously acquired in 10 different trials. The Blue-ROV was intended to move at a speed of 0.5 m/s. However, due to operational deviations, the actual speed varies between 0.3 and 0.6 m/s. Each trial in the data collection experiment took between 5 to 15 min to complete, with the traveling distance ranging approximately from 40 to 450 m.

Figure 9. Pool trial experiment. The figure illustrates various stages of the experimental setup and trials: (a) shows the experimental platform used for conducting the trials, (b,c) highlight the vent models, (d) display the obstacles, (e) depicts the experimental pool where the tests were conducted, (f) represents the joint data acquisition process during the experiments, (g) shows a trial conducted in a bright environment, and (h) presents a trial conducted under dark conditions.
A simulated hydrothermal area environment was recreated to evaluate the performance of the proposed OAF-IPD in SLAM applications. Using videos and images from actual hydrothermal areas as references, scaled and constructed detailed models of hydrothermal deposits were produced. These models were meticulously crafted to mimic the texture, color, and reflective properties of real deposits under various underwater lighting conditions. A comparison between a real hydrothermal smoker and the models under the light source carried by the underwater vehicle is shown in Figure 10.

Additionally, this paper enhanced the complexity of the environment by simulating the frequent visual obstructions encountered during operations, adding four obstacles as shown in Figure 9b. The final simulated hydrothermal vent environment was assembled in a test pool measuring 50 m in length, 30 m in width, and 10 m in depth, with the simulation area covering 30 by 30 m. For the experiments, a Blue-ROV was used as an experimental platform equipped with a Deepsea underwater camera, as well as an M900 multibeam forward-looking sonar. To simulate the various operational scenarios of an AUV in a hydrothermal vent environment, 10 experimental runs were conducted within the constructed environment, including straight-line navigation, pure rotation, comb-like scanning, and circumnavigation around the deposits.

Data for the optical–acoustic model’s mutual supervision training was captured with or without external lightning, as shown in Figure 9g,h.
6.2. Datasets

To ensure a comprehensive assessment of performance across different underwater environments, the data adopted in this paper contain 10 sets of sonar–camera combined data from the pool trial experiment and 18 image sequences from four underwater scenes: the swimming pool, harbor, archaeological, and hydrothermal areas. The distribution of image pairs in each scene is shown in Table 1. Due to the different image sources, the original size of the image is different. It will be resized to $240 \times 320$ and $480 \times 640$ two resolutions before using. The details of the data are as follows:

<table>
<thead>
<tr>
<th>Scene</th>
<th>Sensor Type</th>
<th>Number of Sequences</th>
<th>Number of Data Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swimming pool</td>
<td>Camera</td>
<td>3</td>
<td>536</td>
</tr>
<tr>
<td>Harbor</td>
<td>Camera</td>
<td>3</td>
<td>6225</td>
</tr>
<tr>
<td>Archaeological</td>
<td>Camera</td>
<td>3</td>
<td>3470</td>
</tr>
<tr>
<td>Hydrothermal area</td>
<td>Camera</td>
<td>9</td>
<td>729</td>
</tr>
<tr>
<td>Experimental pool</td>
<td>camera and sonar</td>
<td>10</td>
<td>15,185</td>
</tr>
<tr>
<td>Total</td>
<td>-</td>
<td>28</td>
<td>19,920</td>
</tr>
</tbody>
</table>

Datasets used in this paper is shown in Table 2. The harbor and archaeological data are from the AQUALOC dataset [30], a publicly available underwater dataset gathered near the seabed using ROV. The hydrothermal fields data were sourced from open-access data [31–33]. The swimming pool data and the camera–sonar combined data were collected by the research team of this paper. The dataset collected in the experimental pool simulating a hydrothermal vent environment is named the OA combined dataset.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Data Composition</th>
<th>Data Source</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underwater images dataset</td>
<td>Image sequences</td>
<td>AQUALOC Hydrothermal area data</td>
<td>Camera encoder pretrain IPD evaluation</td>
</tr>
<tr>
<td>Sonar dataset</td>
<td>Sonar data</td>
<td>Pool experiment</td>
<td>Sonar encoder pretrain</td>
</tr>
<tr>
<td>Non-rigid feature dataset</td>
<td>Image sequences</td>
<td>Hydrothermal area data</td>
<td>Non-rigid feature removal training framework</td>
</tr>
<tr>
<td>OA combined dataset</td>
<td>Camera–sonar combined data</td>
<td>Pool experiment</td>
<td>Mutually supervised training SLAM evaluation</td>
</tr>
</tbody>
</table>

6.3. Training Details

The proposed network was trained using PyTorch 1.5.0 [34].

Branch A applied a random homography transformation that included scaling, rotation, and perspective changes, sampled uniformly within specified limits. Additionally, non-spatial augmentations such as noise, blur, and brightness modifications were applied. The maximum distance between corresponding points $\varepsilon_c$ was set to 6. The parameters used for ground truth depth calculation were $\varepsilon_\theta = 0.1$ and detection distance $d_{\text{max}} = 20$. A step size of $\varepsilon_r = 20$ was used during the non-rigid training. A minimum distance of $\varepsilon_f = 4$ was adapted for correspondence determination.

Descriptor loss weights used by both UnSuperPoint and SuperPoint were employed, with a positive margin of $m_p = 1$, a negative margin of $m_n = 0.2$, and a balancing factor of $\lambda_{\text{decu}} = 250$, $\lambda_{\text{son}} = 0.7$, and $\lambda_{\text{dep}} = 0.5$.

The broad and largely unexplored search space for optimal weight estimation for the loss terms in Equation (2) required coarse adjustments for each new term. The selected
weights were 1 for $\alpha_{\text{depth}}$, 1 for $\alpha_p$, 2 for $\alpha_s$, 0.1 for $\alpha_{\text{depth}}$, 0.001 for $\alpha_{\text{desc}}$, and 1 for $\alpha_{\text{rpt}}$. $\alpha_{\text{rig}}$ was set to 1 only when a non-rigid feature dataset was used for specialized training; otherwise, it is set to 0. Our method training consisted of five steps, and the specific steps and detailed parameters are as follows.

(1) Transfer Learning for Camera Data Encoder in Underwater Environments:

This step involved applying transfer learning techniques to adapt a pre-trained camera encoder for underwater environments. The dataset used comprises underwater images with resolutions of $480 \times 640$. The optimizer chosen for this task was Adam, with an initial learning rate of $1 \times 10^{-4}$. The training was conducted in batches of 16 over 10 epochs.

(2) Training for Non-Rigid Interest Points Removal:

This step focused on training the model to remove non-rigid feature points, ensuring the model emphasized rigid structures. The training parameters included a custom loss function based on feature point matching and the use of the SGD optimizer [35]. The learning rate was set at $1 \times 10^{-3}$, with a batch size of 16, and training spans 10 epochs;

(3) Pretraining for Sonar Data Encoder:

In this step, a sonar data encoder was pretrained to effectively handle sonar data. The dataset consisted of sonar data from an experimental pool, with an image resolution of $480 \times 640$. The optimizer used was Adam, with an initial learning rate of $5 \times 10^{-4}$. The batch size for training is 32, and the training is conducted over 20 epochs;

(4) Mutually Supervised Training:

This step involved utilizing mutually supervised learning techniques to train the model using both the camera and sonar data. The dataset was a combination of the camera and sonar data. The optimizer selected was Adam, with a learning rate of $1 \times 10^{-4}$ and a batch size of 16. The training was carried out over 10 epochs for camera-supervised and another 10 for sonar-supervised, employing the joint loss function that incorporates both camera and sonar data;

(5) Final Training for the Entire Model:

The final step involved conducting comprehensive training of the entire model to ensure optimal integration and performance with both the camera and sonar data. The dataset used was a combined dataset of camera and sonar data. The optimizer chosen was Adam, with a learning rate of $1 \times 10^{-5}$ and a batch size of 16. The training process spans 20 epochs, utilizing the joint loss function.

These steps provide a comprehensive process from the initial transfer learning to the final training of the model, ensuring high efficiency and accuracy in underwater environments.

6.4. Results

In this section, the OAF-IPD is compared with other interest point detectors. For SIFT, SURF, and ORB, we used the implementations provided by OpenCV (v3.4.6dev) [14–16]. For both SuperPoint and UnsuperPoint, the author-released GitHub versions were adopted [36,37], and they were trained on the same dataset as the OAF-IPD using the original training parameters with a batch size of 16 and over 20 epochs. SuperPoint, UnsuperPoint and the OAF-IPD were evaluated on GPU (GeForce 1080Ti: manufactured by Micro-Star International, located in New Taipei City, Taiwan), and the remaining detectors were evaluated on a CPU (Intel i7-8700K: manufactured by Intel Corporation, headquartered in Santa Clara, CA, USA).

Separate evaluations have been conducted on the datasets collected in the controlled simulated environment as well as those from the real underwater environment. However, due to the absence of sonar data in the real environment, only the optical data processing component of the OAF-IPD was evaluated. The dataset from the controlled environment was used to evaluate the complete the OAF-IPD system alongside other IPD methods.
6.4.1. Metix

The metrics utilized in this paper are the same as SuperPoint and UnsuperPoint use to evaluate interest point positions through repeatability rates and localization error [19,20]. The overall detector (including the score, position, and descriptors) was also assessed within a homography estimation framework by measuring the matching score and homography accuracy. Below is a brief description of each metric.

Repeatability score (RS): the RS evaluates interest point quality by calculating the ratio of points observed by both viewpoints to the total number [38]. For planar scenes, correspondences between the two camera views are mapped using a homography. Points correspond if their distance is below a defined threshold, \( \rho \). The RS includes only points within the overlapping region of both views and averages the repeatability from each camera’s perspective.

Localization error (LE): the LE represents the average pixel distance between the corresponding points, only including pairs below distance \( \rho \). Similarly to the RS, the LE averages the error calculated in both camera views.

The homography estimation procedure matched descriptors from the images by using nearest neighbor (brute force) matching with cross-check. The homography is estimated with RANSAC [39] via OpenCV, resulting in a homography matrix and a filtered set of matches adhering to the estimated homography.

Matching score (MS): the MS is the ratio of correct matches to all points within the shared view. A correct match consists of two nearest neighbor points in the descriptor space, separated by a pixel distance less than \( \rho \) after transformation to the same view using the ground truth homography.

Homography accuracy (HA): the HA is determined by the homography error (HE), the mean distance between target image corners transformed by ground truth homography \( H_{gt} \) and estimated homography \( H_{est} \). Illustrated in Figure 11, dashed lines represent the distances between corners transformed by \( H_{gt} \) and \( H_{est} \). The HA is the ratio of correctly estimated homographies to the total number, with correctness determined by the HE below a defined tolerance error \( \varepsilon \). The HA is measured at multiple tolerance values.

![Homography error](image.png)

**Figure 11.** Homography error (HE) is defined as the mean distance between the corners of a target image after being transformed by two different homographies: the ground truth homography \( H_{gt} \) and the estimated homography \( H_{est} \). This error \( d_{HE} \) is visually represented by black dashed lines.

6.4.2. Detector Evaluation

Table 3 shows a comparison of repeatability and localization error in a real underwater environment dataset without sonar data. Interest point indicators were calculated...
using NMS. The OAF-IPD (optical only) not only had high repeatability (approximate to SuperPoint in low resolution and UnsuperPoint in high resolution), but also had lowest localization error in the underwater environment. Even without using the sonar data processing component, the OAF-IPD still achieved good performance.

Table 3. Repeatability (higher is better) and localization error (lower is better) for detectors with 240 × 320 and 480 × 640 resolution on real underwater environment dataset. Underlined numbers indicate the best performance in the corresponding evaluation metrics (the same applies to the following tables).

<table>
<thead>
<tr>
<th>Detectors</th>
<th>Repeatability</th>
<th>Localization error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>240 × 320</td>
<td>480 × 640</td>
</tr>
<tr>
<td>ORB</td>
<td>0.527</td>
<td>0.543</td>
</tr>
<tr>
<td>SURF</td>
<td>0.477</td>
<td>0.454</td>
</tr>
<tr>
<td>SIFT</td>
<td>0.455</td>
<td>0.429</td>
</tr>
<tr>
<td>SuperPoint</td>
<td><strong>0.644</strong></td>
<td>0.581</td>
</tr>
<tr>
<td>UnsuperPoint</td>
<td>0.637</td>
<td><strong>0.613</strong></td>
</tr>
<tr>
<td>OAF-IPD (optical only)</td>
<td>0.640</td>
<td>0.603</td>
</tr>
</tbody>
</table>

The OAF-IPD (optical only) and UnsuperPoint have similar performance in terms of matching scores and single responsiveness estimation with high tolerance error. Both perform better in underwater images in environments, and accurately detect valid interest points in underwater images in a swimming pool environment. Additionally, the distribution of interest points is relatively uniform, which is better than the other IPD methods.

Table 4 shows a comparison of repeatability and localization error in a controlled environment dataset with sonar data. We calculated interest point indicators using NMS. The OAF-IPD not only has the highest repeatability but also has the lowest localization error. The data collected in the dark environment under the experimental pool had a significant impact on the other IPD methods. However, supported by sonar data, the OAF-IPD achieved the best performance through its multi-modal fusion structure.

Table 4. Repeatability (higher is better) and localization error (lower is better) for detectors with 240 × 320 and 480 × 640 resolution on the OA combined dataset.

<table>
<thead>
<tr>
<th>Detectors</th>
<th>Repeatability</th>
<th>Localization Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>240 × 320</td>
<td>480 × 640</td>
</tr>
<tr>
<td>ORB</td>
<td>0.510</td>
<td>0.525</td>
</tr>
<tr>
<td>SURF</td>
<td>0.460</td>
<td>0.435</td>
</tr>
<tr>
<td>SIFT</td>
<td>0.440</td>
<td>0.415</td>
</tr>
<tr>
<td>SuperPoint</td>
<td>0.625</td>
<td>0.560</td>
</tr>
<tr>
<td>UnsuperPoint</td>
<td>0.620</td>
<td>0.595</td>
</tr>
<tr>
<td>OAF-IPD</td>
<td><strong>0.649</strong></td>
<td><strong>0.623</strong></td>
</tr>
</tbody>
</table>

6.4.3. Descriptor Evaluation

Table 5 gives the matching scores and homography estimates on the real underwater environment dataset for tolerance thresholds of 1, 3 and 5, respectively. The OAF-IPD (optical only) and UnsuperPoint have the highest match score in different resolutions, which is the best method for homography estimation, while other methods show large errors. SIFT and Superpoint achieve the highest homography estimation accuracy at a tolerance of ε = 1 in respective resolutions. Unsuperpoint gets the highest homography estimation accuracy when ε = 5. The OAF-IPD performs best across other tolerance values.
Table 5. Homography estimation and matching score of detectors for low and medium resolution on real underwater environment dataset.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>240 × 320 300 Points</th>
<th>480 × 640 1000 Points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HE ( \varepsilon = 1 )</td>
<td>HE ( \varepsilon = 3 )</td>
</tr>
<tr>
<td>ORB</td>
<td>0.142</td>
<td>0.456</td>
</tr>
<tr>
<td>SURF</td>
<td>0.404</td>
<td>0.712</td>
</tr>
<tr>
<td>SIFT</td>
<td>0.606</td>
<td>0.842</td>
</tr>
<tr>
<td>SuperPoint</td>
<td>0.487</td>
<td>0.840</td>
</tr>
<tr>
<td>UnsuperPoint</td>
<td>0.530</td>
<td>0.839</td>
</tr>
<tr>
<td>OAF-IPD (optical only)</td>
<td>0.557</td>
<td>0.854</td>
</tr>
</tbody>
</table>

As illustrated in Figure 12, the OAF-IPD (optical only) results in a more dispersed distribution of feature points, which is advantageous for subsequent SLAM backend processing.

Figure 12. Underwater image representation of harbor environment. The above is the result of UnsuperPoint, while the following are the results of the OAF-IPD (optical only). The different colored lines represent different point pairs (the same applies to the subsequent images).

Figure 13 shows the performance of the OAF-IPD (optical only) and UnsuperPoint in the hydrothermal area environment. Both methods perform well on clear, bright, and texture-rich images. However, in the blurry and feature-ambiguous images shown in Figure 13a, the interest points output by UnsuperPoint are fewer and matching errors occur more frequently. In contrast, the OAF-IPD is still able to output sufficient interest points that are repeatably detected and successfully matched.

Furthermore, the OAF-IPD (optical only) significantly outperforms UnsuperPoint when dealing with interference from non-rigid objects. As shown in Figure 13b, many feature points detected by UnsuperPoint are invalid and located on non-rigid objects, whereas all feature points detected by the OAF-IPD are situated on rigid objects.

Table 6 gives the matching scores and homography estimates on the OA combined dataset. The OAF-IPD had the highest homography estimation and match score in different resolutions. Leveraging the training architecture that integrates acoustic and optical visibility, the OAF-IPD demonstrated superior performance in the challenging low-light conditions of the simulated underwater environment. In contrast, when applied to the real underwater environment dataset with better lighting conditions, the HE and MS metrics for other IPD methods showed a noticeable decline.
As illustrated in Figure 12, the OAF-IPD (optical only) results in a more dispersed distribution of feature points, which is advantageous for subsequent SLAM backend processing.

Figure 12. Underwater image representation of harbor environment. The above is the result of UnsuperPoint, while the following are the results of the OAF-IPD (optical only). The different colored lines represent different point pairs (the same applies to the subsequent images).

Figure 13 shows the performance of the OAF-IPD (optical only) and UnsuperPoint in the hydrothermal area environment. Both methods perform well on clear, bright, and texture-rich images. However, in the blurry and feature-ambiguous images shown in Figure 13a, the interest points output by UnsuperPoint are fewer and matching errors occur more frequently. In contrast, the OAF-IPD is still able to output sufficient interest points that are repeatably detected and successfully matched.

Furthermore, the OAF-IPD (optical only) significantly outperforms UnsuperPoint when dealing with interference from non-rigid objects. As shown in Figure 13b, many feature points detected by UnsuperPoint are invalid and located on non-rigid objects, whereas all feature points detected by the OAF-IPD are situated on rigid objects.

(a)  
(b)

Table 6. Homography estimation and matching score of detectors for low and medium resolution on the OA combined dataset.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>240 × 320 300 Points</th>
<th>480 × 640 1000 Points</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>240 × 320 300 Points</td>
<td>480 × 640 1000 Points</td>
<td>ε = 1</td>
<td>ε = 3</td>
<td>ε = 5</td>
<td>ε = 1</td>
</tr>
<tr>
<td>ORB</td>
<td>0.120 0.418 0.522 0.211 0.258 0.563 0.657 0.208</td>
<td>0.381 0.675 0.706 0.260 0.412 0.696 0.759 0.206</td>
<td>0.433 0.704 0.831 0.281 0.481 0.778 0.825 0.247</td>
<td>0.465 0.802 0.837 0.303 0.487 0.775 0.859 0.255</td>
<td>0.507 0.801 0.866 0.386 0.461 0.701 0.862 0.302</td>
<td>0.562 0.859 0.923 0.417 0.512 0.857 0.911 0.397</td>
</tr>
</tbody>
</table>

The OAF-IPD can achieve better repeatability, lower localization error, more accurate homography estimation with low tolerance error, and higher matching scores in more complex underwater images, as evidenced by the more even distribution of feature points and the greater number of correctly matched pairs. As shown in Figure 14, in dark environments, the integration of sonar data significantly enhances the success rate of interest point matching.
Figure 13. Image representation of the hydrothermal area environment. The above is the result of UnsuperPoint, while the following are the results of the OAF-IPD. (a) shows that the OAF-IPD method can obtain more matchable feature points in blurred hydrothermal area images, with a higher accuracy in descriptor matching. (b) illustrates the effect of the non-rigid feature filter. In this figure, the blue box highlights features from rigid objects that are necessary for SLAM, while the orange box shows features from non-rigid objects, which are detrimental to the SLAM system.

Table 6 gives the matching scores and homography estimates on the OA combined dataset. The OAF-IPD had the highest homography estimation and match score in different resolutions. Leveraging the training architecture that integrates acoustic and optical visibility, the OAF-IPD demonstrated superior performance in the challenging low-light conditions of the simulated underwater environment. In contrast, when applied to the real underwater environment dataset with better lighting conditions, the HE and MS metrics for other IPD methods showed a noticeable decline.

Table 6. Homography estimation and matching score of detectors for low and medium resolution on the OA combined dataset.

<table>
<thead>
<tr>
<th>Description</th>
<th>HE</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ε = 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ε = 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ε = 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORB</td>
<td>0.120</td>
<td>0.418</td>
</tr>
<tr>
<td>SURF</td>
<td>0.381</td>
<td>0.675</td>
</tr>
<tr>
<td>SIFT</td>
<td>0.433</td>
<td>0.704</td>
</tr>
<tr>
<td>SuperPoint</td>
<td>0.465</td>
<td>0.802</td>
</tr>
<tr>
<td>UnsuperPoint</td>
<td>0.507</td>
<td>0.801</td>
</tr>
<tr>
<td>OAF-IPD</td>
<td>0.562</td>
<td>0.859</td>
</tr>
</tbody>
</table>

The OAF-IPD can achieve better repeatability, lower localization error, more accurate homography estimation with low tolerance error, and higher matching scores in more complex underwater images, as evidenced by the more even distribution of feature points and the greater number of correctly matched pairs. As shown in Figure 14, in dark environments, the integration of sonar data significantly enhances the success rate of interest point matching.

Figure 14. Image representation of the simulated hydrothermal environment. The above is the result of UnsuperPoint, while the following are the results of the OAF-IPD.

6.4.4. Localization Accuracy Evaluation

To further evaluate the effectiveness of the proposed method in an SLAM system, this section replaces the ORB features in the ORB-SLAM3 with the proposed OAF-IPD method and compares the localization performance of both systems [40]. The vehicle trajectories generated by COLMAP with feature metric refinement are considered as the ground truth for evaluating the accuracy of location. When adjusted based on the deployment positions and pool reference markers, the accuracy of these trajectories can reach the decimeter level.

Table 7 shows the error of location results. The average error trajectory under additional lighting using the OAF-IPD method is 0.17 m, while for ORB it is 0.2 m. In the last five missions, where the only light source was the vehicle’s own illumination, ORB experienced failures in missions 6 and 7, and initialization failures in missions 8, 9, and 10. In contrast, the OAF-IPD successfully completed all tracking in missions 6 and 7, and although it experienced minor losses in missions 8, 9, and 10, it managed to re-localize. Even under these limited lighting conditions, the OAF-IPD maintained a low positioning error of approximately 0.4 m.

Table 7. Positioning results of pool experiment.

<table>
<thead>
<tr>
<th>No.</th>
<th>Length of Trajectory (m)</th>
<th>Number of Frames</th>
<th>Completion Rate (%)</th>
<th>Average Error in the Completed Trial (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>ORB</td>
<td>OAF-IPD</td>
</tr>
<tr>
<td>1</td>
<td>105</td>
<td>1867</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>240</td>
<td>3421</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>327</td>
<td>3343</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>43</td>
<td>633</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>363</td>
<td>5423</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>87</td>
<td>1242</td>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>118</td>
<td>1747</td>
<td>13</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>148</td>
<td>2379</td>
<td>-</td>
<td>90</td>
</tr>
<tr>
<td>9</td>
<td>292</td>
<td>3090</td>
<td>-</td>
<td>86</td>
</tr>
<tr>
<td>10</td>
<td>449</td>
<td>4453</td>
<td>-</td>
<td>93</td>
</tr>
</tbody>
</table>
In the detection of hydrothermal near-field exploration missions, a typical task involves closely circling the smoker body. Figure 15 presents a comparison of the positioning accuracy between the OAF-IPD and ORB during a circling task conducted in a pool experiment, where both methods successfully achieved trajectory positioning. The red line represents the localization achieved by the OAF-IPD method, the blue line represents the localization by the ORB method, and the green line represents the ground truth.

Figure 15. Three-dimensional surround trajectories for hydrothermal smoker model. The red line indicates localization using the OAF-IPD method, the blue line shows localization via the ORB method, and the green line represents the ground truth.

Figure 16 provides a detailed analysis of errors across three dimensions, highlighting the significant accuracy advantage of the OAF-IPD.

Figure 16. Errors in the X, Y, Z directions, respectively.

7. Conclusions

In this paper, the proposed OAF-IPD method effectively addresses the challenges associated with sensor data fusion in hydrothermal near-sea floor environments. By integrating an FPN-based feature fusion module and a depth module, the system ensures a
7. Conclusions

In this paper, the proposed OAF-IPD method effectively addresses the challenges associated with sensor data fusion in hydrothermal near-seafloor environments. By integrating an FPN-based feature fusion module and a depth module, the system ensures a uniform distribution of interest points in depth, thereby enhancing localization accuracy. The self-supervised training procedure enables effective unsupervised training of the OAF-IPD method. The procedure includes an auto-encoder framework for the sonar data encoder training, a ground truth depth generation framework for the depth module training, and an optical–acoustic mutually supervised framework for fuse module training. After the non-rigid feature removal training framework, the camera data encoder acquires non-rigid feature filtering capabilities, reducing the impact of smoke emitted from active vents in hydrothermal areas on localization accuracy. The OAF-IPD system demonstrates significant potential for improving the robustness and accuracy of AUV self-localization in challenging hydrothermal near-seafloor environments.

There are also some shortcomings in our paper, which will be refined in future studies. The validation of the method was conducted in a controlled, artificial environment, which may have resulted in certain differences when compared to real-world conditions. Implementing a dedicated initialization process for fusion sensors may bring further improvement of positioning accuracy, and the graph optimization method the OAF-IPD adopted for position estimation did not take into account the motion characteristics of AUVs.

Author Contributions: Conceptualization, Y.L.; methodology, Y.L.; software, Y.L. and Z.Z.; validation, Y.X.; writing—original draft preparation, Y.L. and J.L.; writing—review and editing, Y.Z. and L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (51979058).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The AQUALOC dataset used in this study is publicly available and can be accessed at https://github.com/SYSU-CPNTLab/LBL-AQUALOC-Dataset (accessed on 18 July 2024). The dataset includes measurements from LBL, a monocular camera, a pressure sensor, and a low-cost MEMS-IMU, all packaged in Rosbag format. The dataset collected by the authors, which includes measurements of images and sonar data, is available upon request. Interested researchers can contact liuyihuiheu@hrbeu.edu.cn for access. Images of hydrothermal vent sites were sourced from the following online platforms: (1) NOAA Ocean Exploration: NOAA Ocean Exploration; (2) PMEL Earth-Ocean Interactions Program: PMEL Earth-Ocean Interactions Program; and (3) InterRidge Vents Database: InterRidge Vents Database. Please ensure appropriate attribution if redistributing or referencing these visuals. For any specific image access or queries, contact the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.