drones

MDPI

Article

Enhancing Inter-AUV Perception: Adaptive 6-DOF Pose Estimation with Synthetic Images for AUV Swarm Sensing

Qingbo Wei [1,2], Yi Yang [1,*], Xingqun Zhou [1], Zhiqiang Hu [1], Yan Li [1], Chuanzhi Fan [1], Quan Zheng [1] and Zhichao Wang [1]

[1] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; weiqingbo22@mails.ucas.ac.cn (Q.W.); zhouxingqun@sia.cn (X.Z.); hzq@sia.cn (Z.H.); liyan@sia.cn (Y.L.); fanchuanzhi@sia.cn (C.F.); zhengquan@sia.cn (Q.Z.); wangzhichao@sia.cn (Z.W.)

[2] University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: yangyi@sia.cn

Abstract: The capabilities of AUV mutual perception and localization are crucial for the development of AUV swarm systems. We propose the AUV6D model, a synthetic image-based approach to enhance inter-AUV perception through 6D pose estimation. Due to the challenge of acquiring accurate 6D pose data, a dataset of simulated underwater images with precise pose labels was generated using Unity3D. Mask-CycleGAN technology was introduced to transform these simulated images into realistic synthetic images, addressing the scarcity of available underwater data. Furthermore, the Color Intermediate Domain Mapping strategy is proposed to ensure alignment across different image styles at pixel and feature levels, enhancing the adaptability of the pose estimation model. Additionally, the Salient Keypoint Vector Voting Mechanism was developed to improve the accuracy and robustness of underwater pose estimation, enabling precise localization even in the presence of occlusions. The experimental results demonstrated that our AUV6D model achieved millimeter-level localization precision and pose estimation errors within five degrees, showing exceptional performance in complex underwater environments. Navigation experiments with two AUVs further verified the model's reliability for mutual 6D pose estimation. This research provides substantial technical support for more complex and precise collaborative operations for AUV swarms in the future.

Keywords: Autonomous Underwater Vehicles (AUVs); 6D pose estimation; underwater perception; environmental adaptation; synthetic underwater images

Citation: Wei, Q.; Yang, Y.; Zhou, X.; Hu, Z.; Li, Y.; Fan, C.; Zheng, Q.; Wang, Z. Enhancing Inter-AUV Perception: Adaptive 6-DOF Pose Estimation with Synthetic Images for AUV Swarm Sensing. *Drones* **2024**, *8*, 486. https://doi.org/10.3390/drones8090486

Academic Editor: Pablo Rodríguez-Gonzálvez

Received: 5 August 2024
Revised: 9 September 2024
Accepted: 12 September 2024
Published: 14 September 2024

1. Introduction

In GNSS-denied underwater environments, achieving precise situational awareness and positioning presents a critical challenge for Autonomous Underwater Vehicles (AUVs). Accurate perception and localization are vital in multi-vehicle coordinated operations and tight formations to ensure precise task execution and the maintenance of formation integrity. Traditional approaches primarily use acoustic localization techniques [1] or acoustic communication devices [2]. Approaches based on acoustic methods have been extensively studied for various tasks, such as docking [3], tracking [4], long-duration cooperative navigation [2], collision avoidance [5], collaborative detection [6], and the maintenance and reconfiguration of tight formations [7]. However, the precision of acoustic equipment is related to its size, which often makes high-precision acoustic equipment cumbersome [8]. Moreover, acoustic localization usually provides only 3D positional data and has a low update rate [9]. In contrast, vision-based methods offer advantages at a near distance, including precise recognition and real-time localization, while providing 6-DOF pose information [10]. This study proposes the AUV6D model, a vision-based 6D pose estimation framework for AUVs, aimed at enhancing the execution of collaborative operations and ensuring precise formations in dynamic underwater environments.

While significant advances have been made in object pose estimation using single RGB images in terrestrial environments, underwater 6D pose estimation still faces two major challenges. First, unlike detecting 2D bounding boxes [11], pose estimation requires 6D pose data, which cannot be manually labeled. Currently, true object poses are extremely difficult to obtain in underwater environments, and hardly any public dataset exists for AUV visual localization, as demonstrated in Figure 1 [12]. Second, the underwater environment presents special challenges due to lighting variations and extensive light scattering and reflection, far exceeding those in terrestrial settings. Underwater images exhibit background variability and substantial differences in color tones and lighting effects [13]. Despite the existence of excellent underwater image processing [14,15] and domain adaptation techniques [16,17], these techniques face challenges in achieving effective adaptation from the source domain to unknown target domains. There is an urgent need for an underwater pose estimation dataset that can be used for single-scene training and multi-scene application. Such a dataset would enhance the adaptability and robustness of visual pose estimation in complex and variable underwater environments.

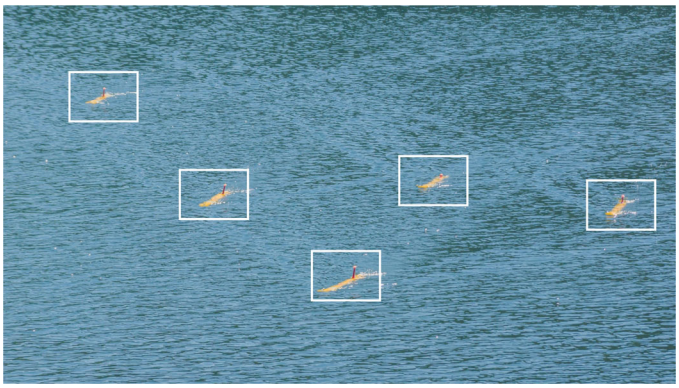

Figure 1. Swarm navigation of "TS-MINI" AUVs. Five TS-MINI AUVs, developed by the Shenyang Institute of Automation, Chinese Academy of Sciences, are performing a surface mission.

To solve these challenges, the AUV6D model for inter-AUV 6D pose estimation is introduced. Synthetic data are used as the source domain, while real underwater images serve as the target domain. The realism of synthetic images is enhanced during training, and feature extraction is standardized during inference. This approach ensures that the model, trained solely on synthetic data, is capable of accurately estimating 6D poses of underwater objects across diverse and complex environments. This is demonstrated through the following contributions.

(1) Generating Synthetic Underwater Images: A comprehensive dataset of underwater simulated images with pose data and object masks was created. A self-supervised CycleGAN with Mask-Cycle Consistency Loss, referred to Mask-CycleGAN, was employed to ensure the realism of 2D object projections. This approach maintains high fidelity and structural consistency before and after the projection of 2D objects. The simulated images were transformed into realistic underwater synthetic images using Mask-CycleGAN. These synthetic images were then used to train the 6D pose estimation model, effectively addressing the scarcity of existing underwater datasets.

(2) Color Intermediate Domain Mapping: During training, synthetic underwater images are first generated by Mask-CycleGAN and then transferred to a defined Color Intermediate Domain (D_{ref}) for training the pose estimation network. This process employs grayscale world white balance based on the B channel of the RGB spectrum to ensure consistent color histogram distributions within the domain. For underwater pose estimation, the integration of color mapping and feature extraction layers at the network's head ensures that various unknown real underwater images align in both pixel and feature dimensions with the training set images. This approach facilitates low-complexity

domain alignment, thereby enhancing the model's adaptability and robustness in diverse underwater environments.

(3) 6D Pose Estimation Network Based on Salient Keypoint Vector Voting: A pose estimation method suitable for underwater environments is proposed, which utilizes a fully convolutional network combined with salient keypoints and directional vectors for voting. This method primarily leverages the shape and structural features of targets, such as centroid points, structural feature points, and bounding boxes on symmetry planes as keypoints. Directional vectors for each pixel are calculated, and voting is performed to determine the positions of the keypoints. This approach enhances the accuracy and robustness of underwater pose estimation and effectively handles occlusions.

(4) Navigation Experiment Validation: To verify the effectiveness and stability of the AUV6D model in practical applications, we utilized the "TS-MINI" AUVs (from Shenyang Institute of Automation, CAS), each equipped with five cameras in the bow section. These cameras cover the upward, downward, leftward, rightward, and forward directions, capturing real-time images of the surrounding AUVs. Experiments were conducted in various water bodies and complex scenarios to evaluate the model's performance. The results indicated that the AUV6D model surpasses existing methods in localization precision and environmental adaptability, successfully locating objects even when targets were occluded. Furthermore, navigation experiments with two "TS-MINI" AUVs demonstrated the capability of the AUV6D model for mutual 6D pose estimation during inter-AUV navigation, with evidence that the relative 6D pose trajectories measured by each AUV closely match.

2. Related Work

In recent years, significant progress has been made in 6D pose estimation technology for AUVs using vision-based methods. However, challenges and limitations remain. This section reviews the related research, including underwater pose estimation using visual markers, deep learning-based pose estimation methods, and advancements in data generation and style transfer.

2.1. Visual Marker-Based Underwater Pose Estimation

Researchers have proposed various methods for underwater pose estimation using visual markers. For instance, one study introduced a method for guiding multiple AUVs using Aruco codes, which achieved high-precision localization [18]. Another study improved visual marker technologies, enhancing the localization accuracy and robustness of underwater robot swarms [19]. Additionally, another study achieved high-precision target localization through monocular vision and low-cost optical beacons [20]. However, visual marker methods face several practical challenges, such as the difficulties of long-term attachment, potential occlusion, and long-distance recognition.

2.2. Deep Learning-Based Pose Estimation

Deep learning-based pose estimation techniques have been widely applied in terrestrial environments, with two primary methods suitable for 6D pose estimation using AUVs: template-based and keypoint-based approaches.

(1) Template-Based Methods: CAD models are used to create a multi-view template database that is then matched against query images. SO-Pose [21] estimates the 6D pose directly using self-occlusion information, while YOLO6D [22] integrates object detection with pose estimation, training on and inferring solely from RGB data. GDR-Net [23] performed 6D pose estimation through a geometry-guided direct regression network. Although these methods are good at handling occlusions and diversity, they heavily rely on object shape, exhibiting poor adaptability to environmental changes, resulting in insufficient robustness in practical applications.

(2) Keypoint-Based Methods: Pose estimation is performed by detecting and matching significant points [24] to enhance the precision of pose estimation by a pixel-level voting network, and HybridPose [25] combined keypoints with symmetry information to perform

well in occluded scenes. The Real-Time Seamless Single Shot 6D Object Pose Prediction method [26] can significantly increase the inference speed. MFPN-6D [27] combines CSPNet and MFPN for efficient single-stage pose estimation. The Keypoint-Graph-Driven Learning Framework [28] can improve the estimation performance in multi-object scenarios, and Keypoint-Guided Efficient Pose Estimation [29] can enhance the robustness of micro-aerial vehicles. However, these methods face challenges in underwater environments due to the scarcity of target texture features and instability of features during navigations.

2.3. Data Generation and Style Transfer

Due to the challenges in acquiring training data from underwater environments, researchers have turned to synthetic data for training purposes. DeepURL [30] generates virtual data using simulation software and employs CycleGAN [31] for image style transfer. Traditional CycleGAN faces issues with the authenticity and controllability of generated images, prompting some studies to propose improvements [32,33]. ROV6D [10] enhanced the diversity and realism of data generation through multi-scene rendering techniques. WaterGAN [34] uses unlabeled real underwater images to generate a substantial volume of synthetic data, supplementing the deficiencies of real data. Although these methods significantly enhance the diversity and realism of synthetic data, the quality and consistency of images generated in unknown environments still require improvement, limiting the model's generalizability and effectiveness in practical applications.

3. Methodology

The AUV6D model was designed to enable 6D pose estimation and improve environmental adaptability among AUVs. The visual system of the 'TS-MINI' AUV and its cooperative localization methods are shown in Figure 2. The architecture of the visual localization system is illustrated in Figure 3. The methodology involves generating synthetic underwater images, applying Color Intermediate Domain Mapping, and constructing the pose estimation model.

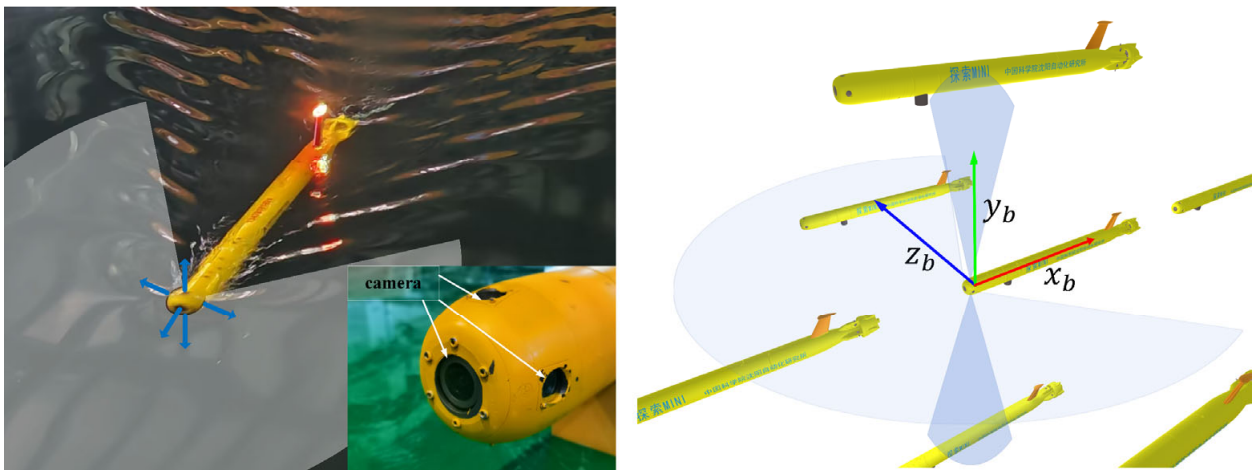

Figure 2. Schematic of the "TS-MINI" AUV and swarm localization. The left image shows a "TS-MINI" AUV sailing at the water surface, with five cameras positioned at the front, top, bottom, left, and right to capture images of the nearby AUVs. The arrows represent the camera views in these directions. The right image illustrates the data collection method used within the swarm and defines the three-dimensional coordinate system.

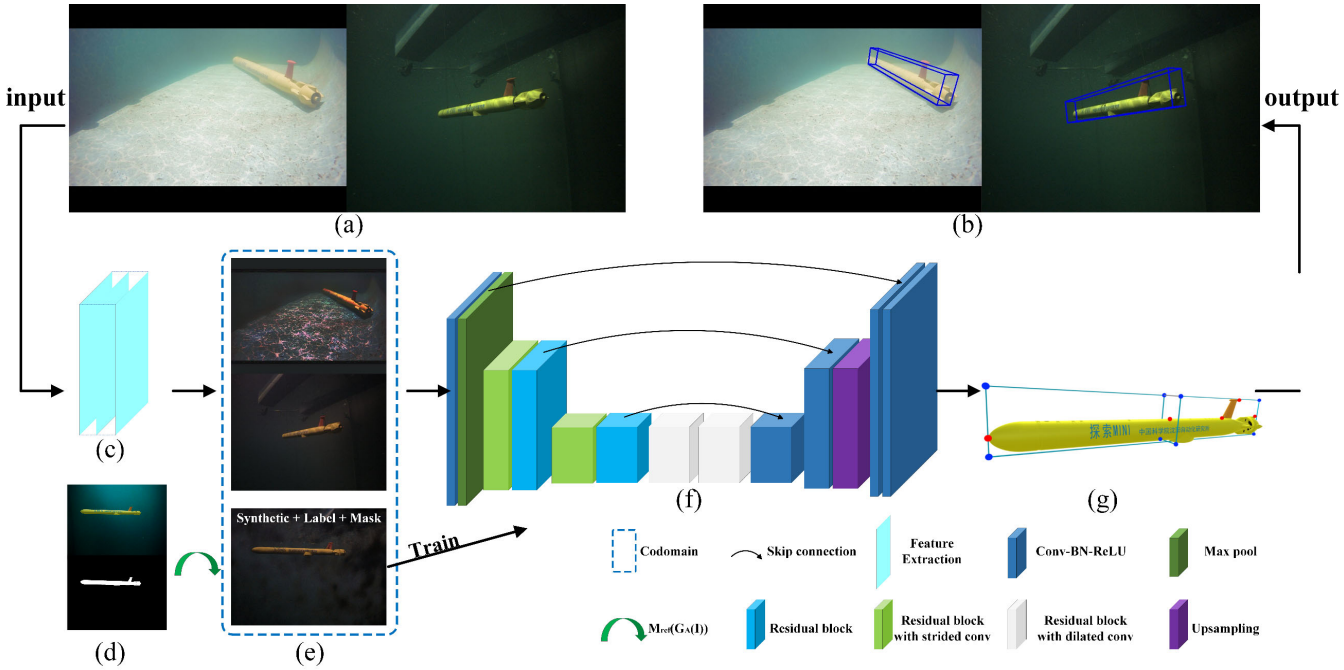

Figure 3. Architecture of the Underwater Autonomous Vehicle visual localization system based on the AUV6D model. (**a**) Images of AUVs collected in underwater environments are used as input data. (**b**) The AUVs' position and orientation are determined by the visual localization algorithm, whose results are indicated by blue boxes. (**c**) Images are mapped to the Color Intermediate Domain and features are extracted. (**d**) Multi-view AUV simulation images, corresponding labeled data, and mask images are generated. (**e**) Realistic images generated by Mask-CycleGAN are labeled and mapped to the Color Intermediate Domain as training data, together with real underwater images that are also mapped to the same domain. $M_{ref}(G_A(I))$ represents the synthetic image transformed into the Color Intermediate Domain from the simulation style. (**f**) The architecture of the pose estimation model includes feature extraction layers, convolutional layers, residual blocks, and upsampling layers, with skip connections for feature transmission. (**g**) Key points of the target are estimated using the extracted features through vector voting. The RT matrix of the target is calculated using the PnP algorithm, which is derived from the pixel coordinates of the keypoints and their corresponding 3D coordinates.

3.1. Generation of Synthetic Underwater Images

Generating realistic underwater images with accurate pose labels is a significant challenge for 6D pose estimation. To solve this, a simulation environment was developed using Unity3D, producing a dataset of simulated images with detailed pose labels and target object masks. However, the complex effects of underwater light propagation cause significant discrepancies between simulated and real images, leading to the poor performance of models trained directly on simulated images.

To address this issue, a Cycle Generative Adversarial Network was used for image transformation. CycleGAN employs a cycle consistency loss, allowing the model to map between simulated and real underwater images. However, the issues of authenticity and controllability still remain [10,32,33]. Therefore, Mask-Cycle Consistency Loss was introduced, leading to the development of Mask-CycleGAN, as shown in Figure 4. This method focuses on target object details during generation, ensuring structural and visual consistency and realism in the generated images.

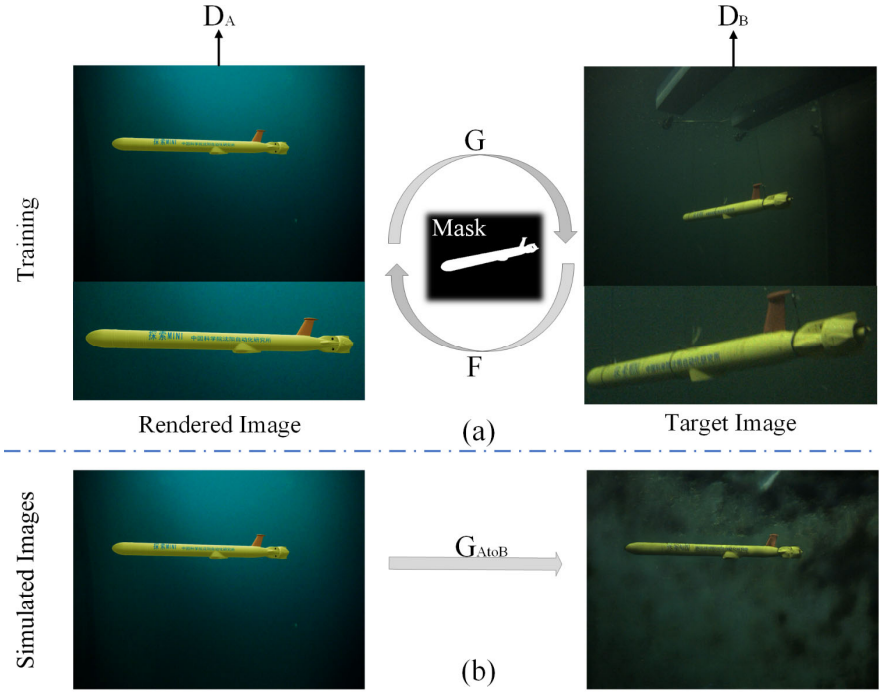

Figure 4. Mask-CycleGAN training and generation process. (**a**) Learning two mapping functions, $G: A \rightarrow B$ and $F: B \rightarrow A$, along with two discriminators, D_A and D_B. The cycle consistency loss incorporates Mask-Cycle Consistency Loss. (**b**) Using only generator G for image transfer, simulated images are converted into the target style images.

During training, mask images are introduced to make the generator focus more on the details within the target object structures. The loss function of Mask-CycleGAN includes Generative Adversarial Loss and Cycle Consistency Loss, defined as follows:

$$\mathscr{L}_{\mathscr{GAN}}(G, D_B, A, B) = E_{b \sim p_{data}(b)}[\ln D_B(b)] + E_{a \sim p_{data}(a)}[\ln(1 - D_B(G(a)))], \quad (1)$$

$$\mathscr{L}_{cyc}(G, F) = E_{a \sim p_{data}(a)}[||F(G(a)) - a||_1] + E_{b \sim p_{data}(b)}[||G(F(b)) - b||_1], \quad (2)$$

Here, E denotes the expectation, and $p_{data}(b)$ and $p_{data}(a)$ represent the data distributions of the target and source domains, respectively. G is the generator, F is the reverse generator, D_B is the discriminator, and A and B are the source and target domain images, respectively. $||\cdot||_1$ denotes the L1 norm.

The introduced Mask-Cycle Consistency Loss is defined as

$$\mathscr{L}_{mask-cyc} = \big|\big|M \odot (A - G(F(A)))\big|\big|_1, \quad (3)$$

where M is the mask image generated from simulation. $G(F(A))$ is the image generated from source domain A to the target domain B and back to the source domain A. $\odot$ denotes element-wise multiplication.

The total loss function is

$$\begin{array}{c} \mathscr{L}_{\mathscr{G},\mathscr{F},\mathscr{D}_{\mathscr{A}},\mathscr{D}_{\mathscr{B}}} = \mathscr{L}_{\mathscr{GAN}}(G, D_B, A, B) + \mathscr{L}_{\mathscr{GAN}}(F, D_A, B, A) \\ +\lambda_{mask-cyc}\left(\lambda_{cyc}\mathscr{L}_{cyc}(G, F) + \lambda_{mask}\mathscr{L}_{mask-cyc}\right) \end{array}, \quad (4)$$

where $\lambda_{mask-cyc}$, λ_{mask}, and λ_{cyc} are weight parameters used to balance the different loss components.

The objective of Mask-CycleGAN is to learn two cycle mapping functions, $G: A \rightarrow B$ and $F: B \rightarrow A$, along with two discriminators, D_A and D_B. By incorporating Mask-Cycle Consistency Loss, the differences in target regions between the generated and original

images are focused on, ensuring consistency and realism in structure and appearance, and facilitating the transformation of simulated images into realistic underwater images.

Experimental Validation: To verify the effectiveness of the Mask-Cycle Consistency Loss, a comparison was conducted between the style transfer results of CycleGAN and Mask-CycleGAN using simulated images, as shown in Figure 5:

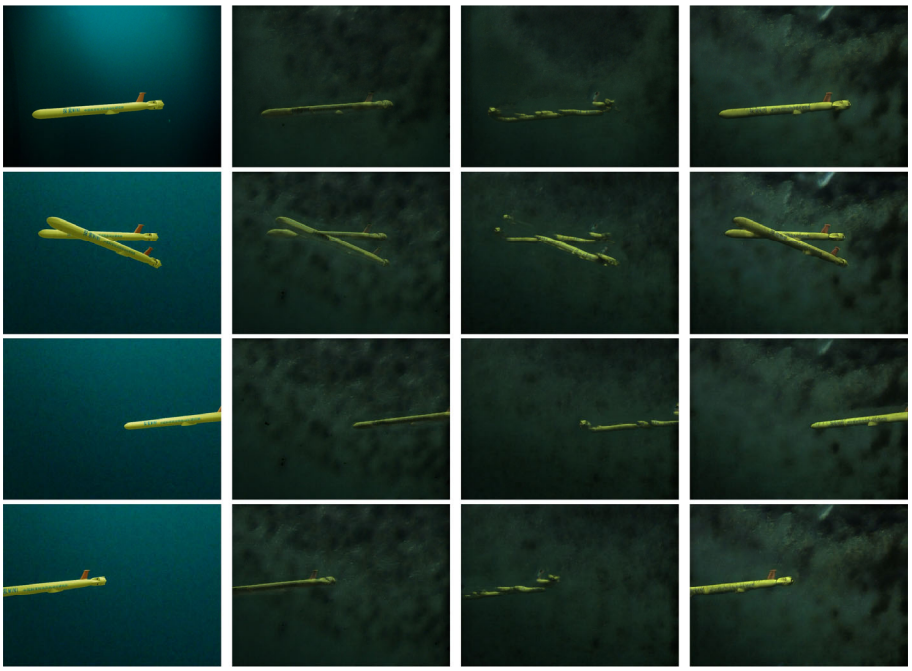

Figure 5. Comparison of image generation results. (**a**) The input simulated image. (**b**) Image generated by CycleGAN lacking realism. (**c**) Image generated by CycleGAN with insufficient structural controllability. (**d**) Image generated by Mask-CycleGAN.

The experimental results demonstrate that Mask-Cycle Consistency Loss, by integrating generative adversarial loss, cycle consistency loss, and mask consistency loss, not only enhances the realism of the generated images but also ensures the consistency of object poses before and after style transfer. The inclusion of Mask-Cycle Consistency Loss significantly improved both the detail realism and pose consistency of the generated images, confirming its efficacy in image generation. This provides high-quality data support for the subsequent training of pose estimation models.

3.2. Image Style Alignment Based on Color Intermediate Domain

To improve the underwater pose estimation model's environmental adaptability, a Color Intermediate Domain Mapping strategy was proposed. Traditional image style transfer techniques produce images with a uniform style, resulting in poor adaptability across varying underwater environments. The Color Intermediate Domain Mapping strategy, by standardizing color histograms and adjusting the white balance based on the blue channel, mitigates color discrepancies, and enhances consistency at the pixel and feature levels.

3.2.1. Definition of the Color Intermediate Domain

The Color Intermediate Domain is established through the following principles.

(1) Uniformity of Channel Color Distribution: Uniformity is achieved through histogram matching, ensuring consistency in color distribution across different images. This method involves adjusting each channel's histograms to match the average histogram of a reference image set, eliminating color discrepancies.

(2) Overall Style Normalization: Based on the theory of grayscale white balance, images are brought to a uniform color balance. This theory assumes that the average of all colors in a natural scene should be neutral (gray), meaning the means of the red, green, and blue channels are equal. White balance adjustments correct color deviations caused by lighting changes, resulting in a more consistent overall image style.

(3) Blue Channel as the Benchmark: The blue channel is chosen for adjustment. Due to its shorter wavelength, blue light has the strongest penetration ability and the least absorption, leading to higher intensity values in underwater environments. It remains more stable in deep water or areas minimally affected by natural light, making it an ideal benchmark for white balance adjustments.

The specific implementation steps are as follows:

Calculation of the Average Histogram for the Reference Image Set

For the RGB channels of the image set in the target style domain of Mask-CycleGAN $\{I_1, I_2, \ldots, I_n\}$, compute the average histogram for each channel:

$$H_R^{ref} = \frac{1}{n}\sum_{i=1}^{n} H_R^i,\ H_G^{ref} = \frac{1}{n}\sum_{i=1}^{n} H_G^i,\ H_B^{ref} = \frac{1}{n}\sum_{i=1}^{n} H_B^i, \tag{5}$$

White Balance Adjustment Based on the Blue Channel

Adjust the red and green channels based on the blue channel to eliminate color distortions:

$$\text{scale}_R = \frac{\mu_B}{\mu_R},\ \text{scale}_G = \frac{\mu_B}{\mu_G}, \tag{6}$$

$$R'' = R' \cdot \text{scale}_R,\ G'' = G' \cdot \text{scale}_G,\ B'' = B', \tag{7}$$

where R', G', and B' represent the pixel values of the red, green, and blue channels after histogram matching; μ_R, μ_G, and μ_B represent the average values of the red, green, and blue channels; and scale_R and scale_G are the adjustment ratio coefficients for the red and green channels relative to the blue channel.

Finally, the Color Intermediate Domain D_{ref} is defined as

$$D_{ref} = \left\{ I \;\middle|\; H_R(I) = H_R^{ref}, H_G(I) = H_G^{ref}, H_B(I) = H_B^{ref}, \mu_{R''} = \mu_{G''} = \mu_{B''} \right\}, \tag{8}$$

The Color Intermediate Domain ensures image consistency and alignment by matching the RGB channel color histograms to the intermediate domain histograms and adjusting the red and green channels based on the blue channel for white balance.

3.2.2. Generating the Synthetic Image Training Set in the Color Intermediate Domain

The proposed strategy aimed to generate a synthetic image dataset suitable for estimating the 6D pose of targets in unknown underwater environments. The Mask-CycleGAN generator G_A transforms the original images I into realistic underwater images $G_A(I)$. Subsequently, the mapping function M_{ref} is applied to map these realistic underwater images $G_A(I)$ into the Color Intermediate Domain H_{ref}. This mapping function ensures consistency in color and lighting characteristics across all images by matching the RGB channel histograms and adjusting the white balance based on the blue channel. The final synthetic images in the Color Intermediate Domain are given by

$$I_{\text{mapped}} = M_{ref}\Big(G_A(I), H_{ref}\Big), \tag{9}$$

As shown in Figure 6, this strategy harmonizes simulated images and various styles of underwater images into a uniform style. After training the pose estimation network with the synthetic images mapped to the Color Intermediate Domain, real underwater images input into the network are processed through image color mapping and feature extraction layers. This ensures that various real images in unknown underwater environments align

with the style of the training set images in terms of pixel and feature dimensions, thereby enhancing the accuracy of cross-scenario pose estimation.

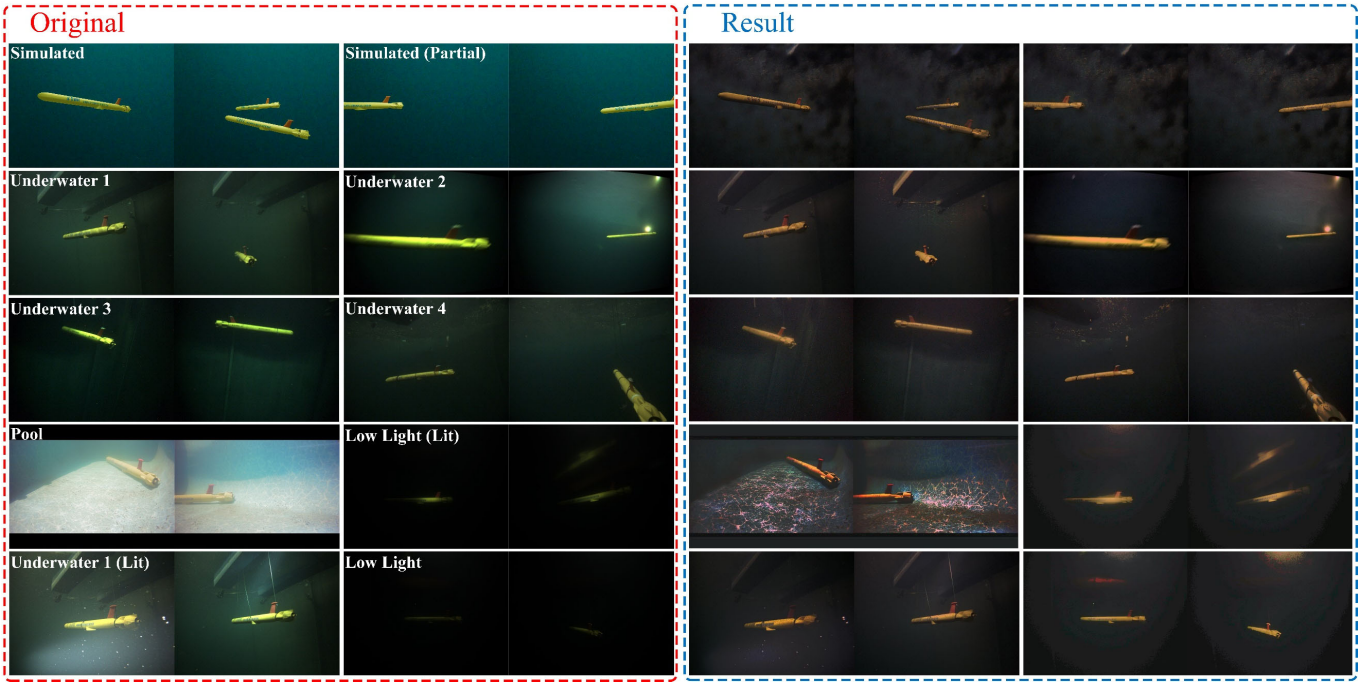

Figure 6. Color mapping of simulated and various real underwater images into the intermediate domain. The left side displays the original images, which include simulated images, images from Water Area 1, Water Area 2, Water Area 3, Water Area 4, a swimming pool, low light conditions, illuminated conditions, and dim lighting conditions. The right side shows the results after mapping to the intermediate domain.

3.3. Pose Estimation Network Based on Salient Keypoint Vector Voting

In non-underwater environments, keypoint-based methods are widely used for six-degrees-of-freedom (6-DoF) pose estimation due to their precision and robustness. However, in underwater settings, the propagation effects of light significantly weaken the target texture features. Additionally, the cylindrical structure of Autonomous Underwater Vehicles (AUVs) lacks distinctive features, which makes the keypoint localization complicated. Furthermore, rotation and changes in angle can create ambiguities in two-dimensional projections, and using external 3D bounding box corners may lead to significant errors due to the AUV's shape. The elongated cylindrical structure of the AUV and its nearly horizontal navigation posture make it likely for its head or tail to exceed the field of view. Potential underwater obstacles further increase localization complexity. Therefore, a pose estimation network based on Salient Keypoint Vector Voting was proposed, incorporating the shape and contour features of the AUV to address these challenges, ultimately enhancing keypoint localization and 6D pose estimation accuracy.

3.3.1. Definition of Keypoints

In 6DoF pose estimation for AUVs, keypoints are selected based on criteria including low ambiguity during rotation and movement, high visibility, and distinctiveness. As shown in Figure 7, 13 keypoints were identified and categorized as follows.

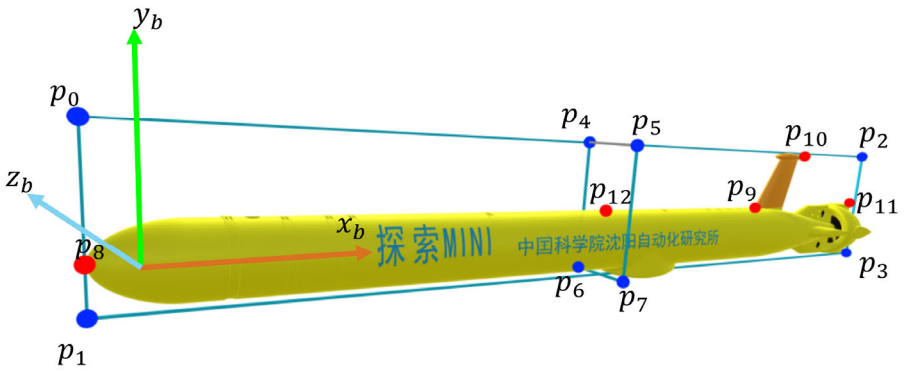

Figure 7. Definition of keypoints. This figure displays the keypoints selected for the 6DoF pose estimation of an AUV, marked from p_0 to p_{12}.

Salient Feature Points: Four salient feature points (p_8, p_9, p_{10}, and p_{11}) were selected on the model's symmetrical xoy plane to reduce ambiguity from different viewpoints.

Centroid: The centroid of the model, p_{12}, was selected.

Extrapolated Rectangle Points on Planes: The four corners of extrapolated rectangles on the model's xoy and yoz plane were chosen.

xoy plane: p_0, p_1, p_2, and p_3.

yoz plane: p_4, p_5, p_6, and p_7.

Given that the AUV exhibits minimal roll during navigation, making the xoz plane projection mostly linear, vertices from this plane were not selected.

To ensure that the keypoints have salient features on the 3D model, the sharpness of each candidate point is computed, which was defined as

$$S(p) = \max_{q \in N(p,r)} \left(\sqrt{\left(\frac{\partial z_q}{\partial x}\right)^2 + \left(\frac{\partial z_q}{\partial y}\right)^2 + \left(\frac{\partial z_q}{\partial z}\right)^2} \right), \tag{10}$$

where $N(p,r)$ denotes the set of points within a radius r of point p, and $\frac{\partial z_q}{\partial x}$, $\frac{\partial z_q}{\partial y}$, and $\frac{\partial z_q}{\partial z}$ represent the gradients of point q in the x, y, and z directions, respectively. The point with the highest sharpness was selected as a salient keypoint.

3.3.2. Vector Voting Model

The AUV6D model utilizes a fully convolutional network architecture, accepting an image with dimensions of H (height) $\times$ W(width) $\times$ 3 (color channels) and outputting a tensor with dimensions of $H \times W \times (K \times 2 + C)$, representing vector fields and class probabilities, where K represents the number of keypoints and C represents the number of object classes. The backbone employs a pre-trained ResNet-18, modified by removing subsequent pooling layers and incorporating dilated convolutions and convolutional layers after reaching feature map dimensions of $H/8 \times W/8$. Unit vectors and class probabilities are derived via 1×1 convolutions.

Semantic segmentation and keypoint detection are concurrently executed. Each pixel emits a semantic label and a directional vector to the 2D keypoints, facilitating the generation of keypoint hypotheses and their localization through a weighted average.

A vector voting approach was adopted to mitigate occlusion issues encountered during AUV navigation. As depicted in Figure 8, directional vectors from each pixel to the keypoints within the target contour are calculated, allowing for accurate localization of occluded keypoints even under high obstruction or partial truncation.

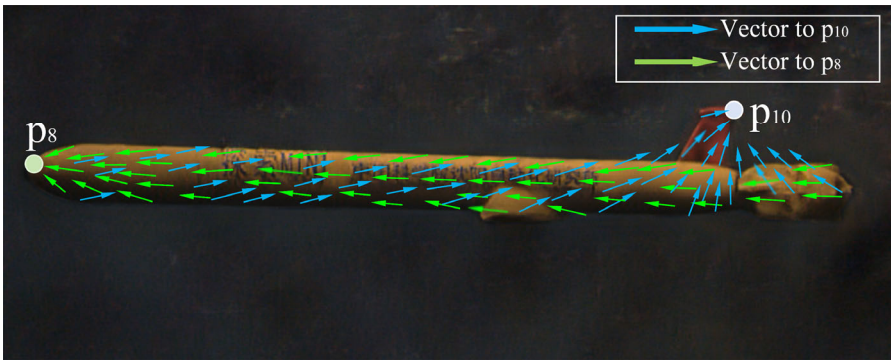

Figure 8. Vector voting diagram. Direction vectors pointing towards keypoints p_8 and p_{10} are illustrated by green and blue arrows, respectively.

The directional vector voting model is mathematically expressed as

$$\boldsymbol{d_i} = \frac{\boldsymbol{p_i} - \boldsymbol{p}}{||\boldsymbol{p_i} - \boldsymbol{p}||}, \tag{11}$$

where $\boldsymbol{d_i}$ signifies the direction vector from pixel position $\boldsymbol{p}$ to keypoint $\boldsymbol{p_i}$, normalized to maintain scale invariance.

The mathematical model for vector voting is defined as

$$V(\boldsymbol{p}, \boldsymbol{p_i}) = \begin{cases} 1, & if\ ||\boldsymbol{d} - \boldsymbol{d_i}|| < \epsilon \\ 0, & \text{otherwise} \end{cases}, \tag{12}$$

where $V(\boldsymbol{p}, \boldsymbol{p_i})$ denotes the voting value of pixel $\boldsymbol{p}$ for keypoint $\boldsymbol{p_i}$, $||\boldsymbol{d} - \boldsymbol{d_i}||$ represents the distance between direction vectors, and ϵ is the voting threshold.

Each pixel casts a vote for all possible keypoint locations based on its directional vector, and the position of each keypoint is determined through weighted averaging:

$$\boldsymbol{p_i} = \frac{\sum_p V(\boldsymbol{p}, \boldsymbol{p_i})\boldsymbol{p}}{\sum_p V(\boldsymbol{p}, \boldsymbol{p_i})}, \tag{13}$$

where $\sum_p V(\boldsymbol{p}, \boldsymbol{p_i})\boldsymbol{p}$ represents the weighted sum of all votes and $\sum_p V(\boldsymbol{p}, \boldsymbol{p_i})$ denotes the total number of votes.

To estimate keypoint positions more precisely, multiple location hypotheses are generated using directional vector voting. These hypotheses' statistical characteristics are then used to calculate the spatial probability distribution of the keypoints. During the RANSAC process, a series of keypoint position hypotheses h_i are generated, each associated with a voting weight w_i. The mean and covariance matrix of keypoint positions are determined by weighted averaging:

$$\mu_k = \frac{\sum_i w_i h_i}{\sum_i w_i}, \tag{14}$$

where $\sum_i w_i h_i$ is the weighted covariance sum of all hypothesis locations and $\sum_i w_i$ is the total weight sum.

$$\Sigma_k = \frac{\sum_i w_i (h_i - \mu_k)(h_i - \mu_k)^T}{\sum_i w_i}, \tag{15}$$

where $\sum_i w_i (h_i - \mu_k)(h_i - \mu_k)^T$ is the weighted covariance sum of all hypothesis locations.

This approach not only yields estimated keypoint positions but also provides statistical information about their spatial distribution for subsequent uncertainty analysis and optimization. The confidence in keypoint selection is determined by evaluating the

consistency of votes, measured by variance; a lower variance indicates more consistent results and more reliable keypoint positions. The confidence formula is

$$C_i = \frac{1}{1 + \sigma_i^2}, \tag{16}$$

where σ_i^2 is the variance of the voting results.

Once keypoint positions are determined, pose estimation is conducted using these points. An uncertainty-driven PnP algorithm is employed, minimizing the projection error of keypoints in the image to estimate the object's 6DoF posture. The mathematical model is expressed as

$$E(R, t) = \sum_i \left| \boldsymbol{p}_i - \pi(R\boldsymbol{P}_i + t) \right|_2, \tag{17}$$

where R represents the object's rotation matrix, t is the translation vector, $\pi(\cdot)$ is the projection function that projects 3D points onto the 2D image plane, $\boldsymbol{P_i}$ represents the 3D keypoints, and $\boldsymbol{p_i}$ represents their projections in the 2D image.

The loss function was designed to optimize the accuracy of keypoint positioning and pose estimation. During training, a multi-task loss function considers both the keypoint localization error and pose estimation error:

$$\mathscr{L} = \lambda_1 \sum_i \left| \boldsymbol{p}_i^{\mathrm{gt}} - \boldsymbol{p}_i \right|_2 + \lambda_2 E(R, t), \tag{18}$$

where $\boldsymbol{p}_i^{\mathrm{gt}}$ is the ground truth keypoint position, and λ_1 and λ_2 are the weight parameters for balancing the keypoint localization and pose estimation errors. Continuous optimization of this loss function, along with data augmentation techniques and an expanded training sample, enhances the model's generalization ability and accuracy.

4. Evaluation of Color Intermediate Domain Mapping Strategy

This experiment aimed to validate the impact of the Color Intermediate Domain Mapping strategy on the similarity between synthetic images and various real underwater images. A comparative experiment was designed to quantify the differences between different training sets (Synthetic) and test sets (Real Images). The assessment included the following categories of synthetic images: directly generated simulated images (Simulated), images generated by CycleGAN lacking realism (CycleGAN IR), images generated by CycleGAN with insufficient controllability (CycleGAN IC), images generated by Mask-CycleGAN (Mask-Cycle), and synthetic images mapped to the Color Intermediate Domain (Intermediate). The real images in the test set originated from five types of water bodies and three additional operational conditions.

Four image quality assessment metrics were employed to quantify image similarity.

(1) Structural Similarity Index (SSIM) [35]: Measures the similarity of two images in terms of brightness, contrast, and structural information. Higher values indicate greater similarity.

$$\mathrm{SSIM}(x, y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}, \tag{19}$$

where μ_x and μ_y are the mean values of images x and y, σ_x^2 and σ_y^2 are the variances of x and y, σ_{xy} is the covariance, and C_1 and C_2 are constants introduced for stability.

(2) Peak Signal-to-Noise Ratio (PSNR) [36]: Based on pixel differences and measures image quality. Higher values indicate better quality.

$$\mathrm{PSNR} = 10 \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right), \tag{20}$$

where MAX is the maximum pixel value and MSE is the mean squared error.

$$\text{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2, \tag{21}$$

where $I(i,j)$ and $K(i,j)$ representing the pixel values of two images, and m and n representing the width and height of the images, respectively.

(3) Gram Matrix Mean Squared Error (Gram MSE) [37]: Used to assess the similarity of image styles. Lower errors indicate more similar styles.

$$G_{ij} = \sum_k F_{ik}F_{jk}, \tag{22}$$

where F represents the image feature matrix.

$$MSE_{\text{loss}} = \frac{1}{c}\sum_{i=1}^{c}\left(G_x(i) - G_y(i)\right)^2, \tag{23}$$

where G_x and G_y are the Gram matrices of images x and y, and c is the number of channels.

(4) Fréchet Inception Distance (FID) [38]: Evaluates the difference in distribution between generated and real images in high-dimensional space. Lower values indicate closer distributions.

$$\text{FID} = |\mu_r - \mu_f|^2 + \text{Tr}\left(\Sigma_r + \Sigma_f - 2\left(\Sigma_r\Sigma_f\right)^{1/2}\right), \tag{24}$$

where μ_r and μ_f are the means of real and generated image features, Σ_r and Σ_f are the covariance matrices of real and generated image features, and Tr denotes the trace of a matrix.

In Table 1, the mean and variance of the different methods for the training set images and eight real-world environmental images under four evaluation metrics are presented. The results indicate that synthetic images mapped to the Color Intermediate Domain achieved the best average results in terms of SSIM, PSNR, Gram matrix mean squared error, and FID. Additionally, the Gram matrix mean squared error, which measures the similarity of image styles, was significantly lower for these images compared to other synthetic images. Except for the SSIM standard deviation, which was slightly higher than that of simulated images, the variance in all the other metrics was the smallest for synthetic images mapped to the Color Intermediate Domain, indicating a higher consistency and stability across different environments.

Table 1. Results of various evaluation metrics for different methods.

Synthetic	SSIM Mean ↑	SSIM Std ↓	PSNR Mean (dB) ↑	PSNR Std (dB) ↓	Gram Mean ↓	Gram Std ↓	FID Mean ↓	FID Std ↓
Simulated	0.3408	0.0736	14.9744	2.7336	0.0018	0.0019	3.4018	0.7189
CycleGAN IR	0.5293	0.1095	16.4843	3.4098	0.0014	0.0021	3.4789	1.0076
CycleGAN IC	0.5694	0.1198	20.0729	5.2012	0.0012	0.0026	2.5584	0.7462
Mask-Cycle	0.5053	0.1090	18.3925	4.3192	0.0012	0.0024	3.2607	0.8832
Intermediate	0.6433	0.0983	22.2661	2.5756	0.0002	0.0005	2.3593	0.6218

Red indicates the optimal value. ↑ indicates higher is better, and ↓ indicates lower is better.

As illustrated in Figure 9, the evaluation trajectory lines for synthetic images mapped to the Color Intermediate Domain were positioned on the outer perimeter of the SSIM and PSNR radar charts, consistently yielding the best results across all environments and closely resembling a regular octagon. In the radar charts for the Gram matrix mean squared error and FID, the trajectory lines were on the inner side, showing the most balanced shape. These results indicate that the Color Intermediate Domain Mapping strategy effectively reduces the gap between synthetic and real images, enhancing consistency and stability at the pixel and semantic feature levels.

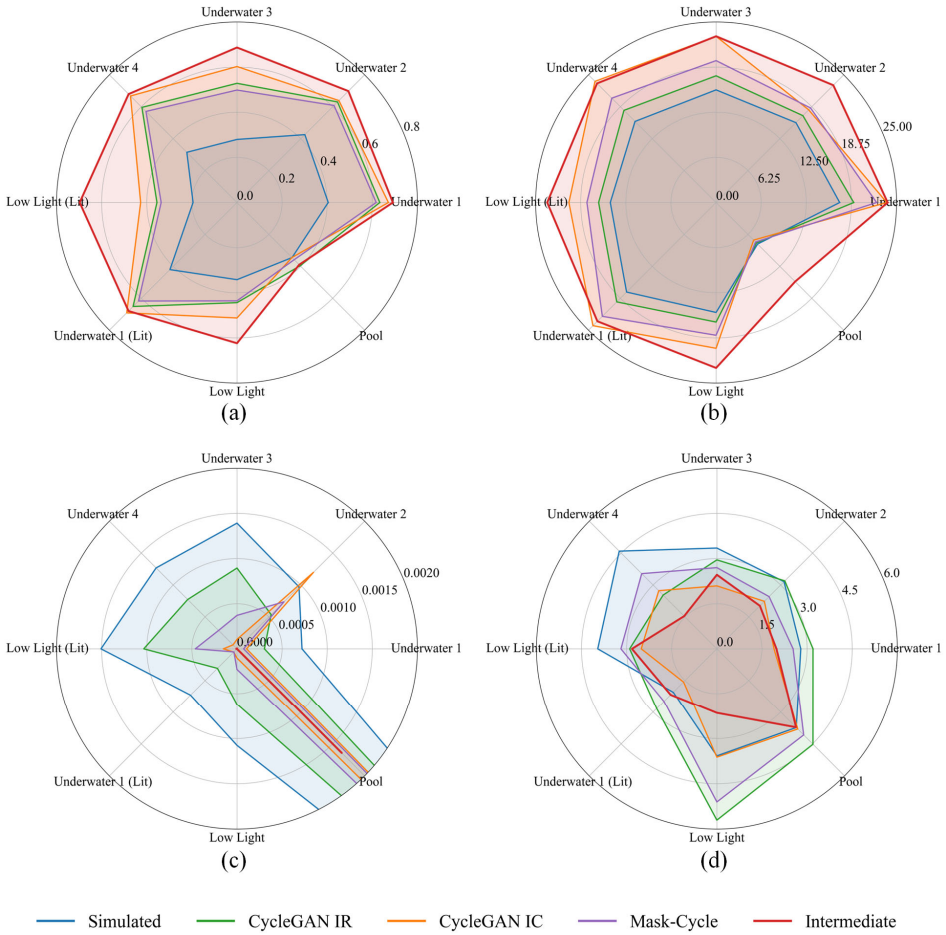

Figure 9. Radar charts of four similarity metrics for five types of synthetic data across eight environments. Each subplot shows the performance of the different methods across various environments in terms of (**a**) Structural Similarity Index (SSIM), (**b**) Peak Signal-to-Noise Ratio (PSNR), (**c**) Gram matrix mean squared error, and (**d**) Fréchet Inception Distance (FID).

Overall, the Color Intermediate Domain Mapping strategy significantly enhanced the similarity between synthetic images and various real underwater images, validating its potential for image style transfer and environmental adaptability applications.

5. Pose Estimation Experiment

5.1. Validation of AUV6D in Dynamic Environments

The objective of this experiment was to evaluate the AUV6D model's performance in estimating 6D poses within dynamic underwater environments. The model was trained exclusively on 10,375 synthetic images mapped to the Color Intermediate Domain, covering the full spectrum of AUV poses encountered during typical navigation. During inference, the model was tested on a diverse set of real underwater images, representing various operational conditions, including different water bodies and lighting scenarios. All images had a resolution of 640 $\times$ 480 pixels.

The model was tested across a range of challenging conditions, including Water Area 1, low-light and supplementary lighting conditions, a swimming pool, lit conditions in Water Area 1, occlusion scenarios, and additional water bodies. Figure 10 illustrates the pose estimation results across these environments, with the AUV's position and orientation represented by blue bounding boxes.

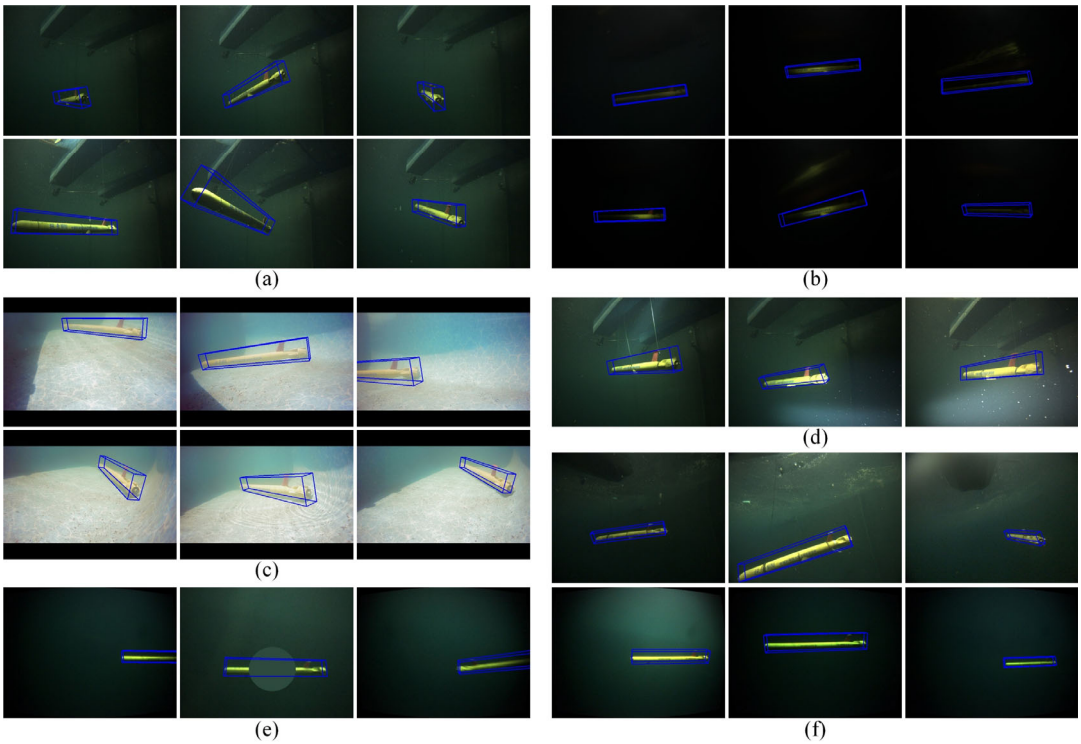

Figure 10. Pose estimation results across various water bodies and conditions. The AUV's position and orientation are delineated with blue bounding boxes. Subplots (**a**–**f**) display the results under the different conditions: (**a**) Water Area 1, (**b**) low-light and supplementary lighting conditions, (**c**) swimming pool, (**d**) lit conditions in Water Area 1, (**e**) various occlusion scenarios, and (**f**) other water bodies.

The findings indicate that the AUV6D model, trained solely on synthetic images from the Color Intermediate Domain, successfully estimated the AUV's pose across diverse environments. These include significantly different image styles, such as swimming pools, low-light conditions, supplementary lighting, and partial occlusions. The results demonstrate the model's robust adaptability to various environmental conditions, confirming its efficacy in dynamic underwater scenarios.

5.2. Evaluation of Environmental Adaptability

To assess the environmental adaptability of the AUV6D model using the Color Intermediate Domain Mapping strategy, four distinct models were trained on different datasets. These datasets included original simulated images, CycleGAN-generated images, Mask-CycleGAN-generated images, and synthetic images mapped to the Color Intermediate Domain (referred to as the "Intermediate" dataset). The "Intermediate" dataset, created through Mask-CycleGAN and subsequently refined via color mapping, was specifically designed to improve adaptability in diverse underwater environments.

For each underwater scenario, 100 images were randomly selected, and the 6D pose estimation results were visually inspected to assess localization success. Table 2 presents the localization success rates of the AUV6D models trained on the four datasets.

Table 2. Localization success rates of AUV6D models trained with different datasets across various water conditions.

	Pose Estimation Success Rate							
	Underwater 1	**Pool**	**Underwater 2**	**Underwater 3**	**Underwater 4**	**Low Light**	**Low Light (Lit)**	**Underwater 1 (Lit)**
Simulated	0	0	0	0	0	0	0	0
CycleGAN	0.23	0	0	0.11	0.04	0	0.01	0.11
Mask-Cycle	0.94	0.02	0.31	0.81	0.19	0.07	0.10	0.76
Intermediate	0.98	0.62	0.94	0.92	0.82	0.75	0.68	0.83

As shown in Table 2, the model trained on simulated images failed to accurately estimate poses in real underwater conditions. The CycleGAN-trained model showed moderate performance in specific environments, such as Underwater 1, but failed in more complex scenarios. In contrast, the Mask-CycleGAN-trained model exhibited higher success rates in the target domain but struggled in other environments. The model trained on the "Intermediate" dataset consistently achieved high success rates across all environments, showcasing the robustness and adaptability of the AUV6D model when employing the Color Intermediate Domain Mapping strategy.

5.3. Evaluation of 6D Pose Estimation

5.3.1. Pose Accuracy Analysis

To assess the 6D pose estimation accuracy of the AUV6D model, a ground truth dataset with 6D pose labels was required. A 50 cm × 50 cm Aruco marker board was installed on the AUV, and the marker's pose was determined using the PnP algorithm. True poses of the AUV within the images were derived using pose calibration techniques, resulting in the dataset depicted in Figure 11. Images were captured at a resolution of 640 × 480 pixels within a distance range of 2 to 5 m. Due to the Aruco marker tied to it, the AUV lost its navigation capability. Therefore, the AUV was suspended and towed underwater to capture these images in this experiment.

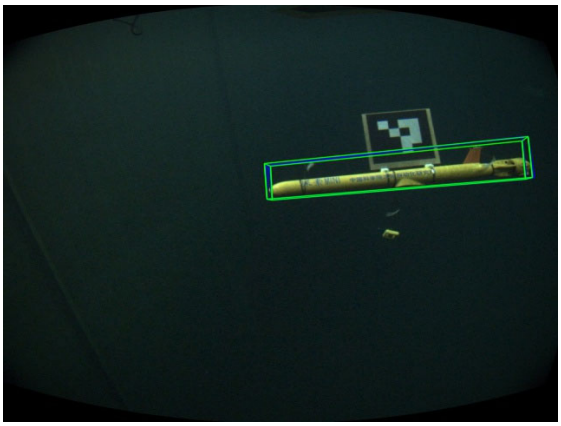

Figure 11. Ground truth dataset. The green box represents the true pose, and the blue box indicates the estimated pose.

Errors in rotation and translation were calculated by comparing the poses estimated by the AUV6D model to those in the ground truth dataset:

$$Error_{rot} = arccos\left(\frac{trace\left(R\hat{R}^T\right) - 1}{2}\right), \tag{25}$$

$$Error_{trans} = \left|t - \hat{t}\right|_2, \tag{26}$$

where R and $\hat{R}$ denote the actual and estimated rotation matrices, and t and $\hat{t}$ denote the actual and estimated translation vectors, respectively.

The errors under all six degrees of freedom were analyzed, and the distributions of rotational and translational deviations are illustrated in Figure 12 using box plots. The box plots revealed that the deviations predominantly ranged as follows: *X*-direction deviations were between 0.01 m and 0.04 m, *Y*-direction deviations were between 0.005 m and 0.025 m, and *Z*-direction deviations were between 0.03 m and 0.20 m; roll errors were between 2.5 degrees and 13 degrees, pitch errors were between 2 degrees and 7 degrees, and yaw errors were between 0.2 degrees and 3 degrees. The mean errors for all six axes are shown in Table 3, with the *X* and *Y* directional errors averaging at the millimeter level, and the pitch and yaw errors were less than 5 degrees. However, the *Z*-direction errors are comparatively higher, with the largest errors in the roll direction, due to the high sensitivity of the *Z*-axis and roll angle calculations in the PnP algorithm. Additionally, the AUV's elongated cylindrical structure leads to less noticeable displacement changes along the *X*-axis, with limited feature changes around the *X*-axis.

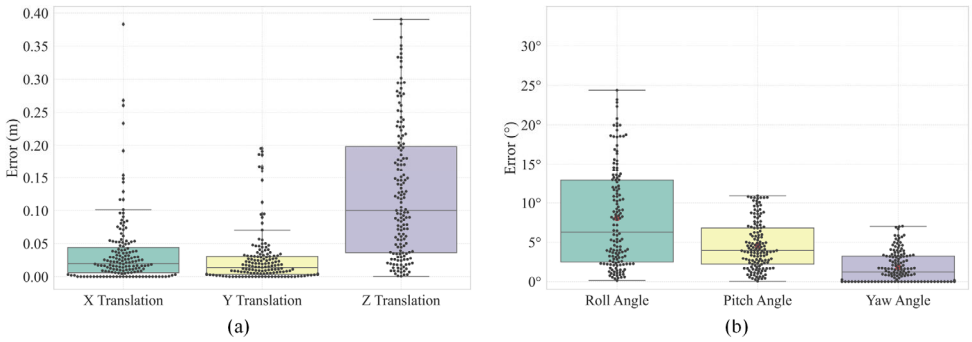

Figure 12. Error box plots for the AUV6D model. (**a**) Translational error distribution across *X*-, *Y*-, and *Z*-directions. (**b**) Euler angle error distribution across roll, pitch, and yaw.

Table 3. Mean errors in six-degrees-of-freedom pose estimation for AUV.

Translation *X* Error (m)	Translation *Y* Error (m)	Translation *Z* Error (m)	Roll Error (m)	Pitch Error (m)	Yaw Error (m)
0.0248	0.0168	0.1099	8.0653	4.5888	1.8232

5.3.2. Comparative Analysis of Methods

The AUV6D model was compared with mainstream methods such as DeepURL [30], PVNet [24], and YOLO6D [26] in terms of localization accuracy and computational efficiency. The experimental dataset comprised 10,375 synthetic images in the Color Intermediate Domain, which was evaluated against the ground truth dataset. The frame rates of all algorithms were tested on an RTX 2080 graphics card, as summarized in Table 4.

Table 4. Comparison of different algorithms in terms of localization accuracy and computational efficiency.

	Translation Error (m)	Orientation Error (°)	ADD	FPS
DEEPURL	0.068	6.77°	57.16%	40
Intermediate + PVNET	0.186	14.55°	53.09%	37
Intermediate + YOLO6D	0.472	20.56	33.49%	54
AUV6D	0.051	4.83	62.63%	38

Red indicates the optimal value.

As shown in Table 4, the AUV6D model outperformed the current mainstream methods in terms of localization accuracy, exhibiting the smallest translation and rotation errors, and achieving the highest ADD score, thus demonstrating its precision advantage. Although its frame rate was slightly lower than that of YOLO6D, it was comparable to that

of DeepURL and PVNet; overall, it displayed a high computational efficiency suitable for high-precision localization and real-time computation in complex underwater environments.

6. Navigation Experiments

The practical application of the AUV6D model in underwater navigation was demonstrated through a series of navigation experiments using "TS-MINI" AUVs. To assess the absolute differences in 6D poses between the estimated and true values, a tow navigation experiment was performed with a Aruco marker board installed on the AUVs. Additionally, to verify the inter-AUV 6D pose estimation capability, two "TS-MINI" AUVs were used for mutual pose estimation during autonomous navigation. The reliability of inter-AUV localization was validated by comparing the mutual estimation results of the two AUVs.

6.1. Tow Navigation Experiment

In this experiment, one AUV equipped with a visual marker was towed along a predetermined trajectory, while another AUV captured footage. The AUV6D model estimated the 6D pose of the towed AUV using the video data, and these estimates were compared against the ground truth 6D poses provided by the visual markers. As depicted in Figure 13, the trajectory estimated by the AUV6D model closely aligned with the actual trajectory, with minimal translational and rotational errors. This validates the model's accuracy in estimating the pose and position of nearby AUVs, demonstrating its capability to provide accurate and rich information for inter-AUV coordination.

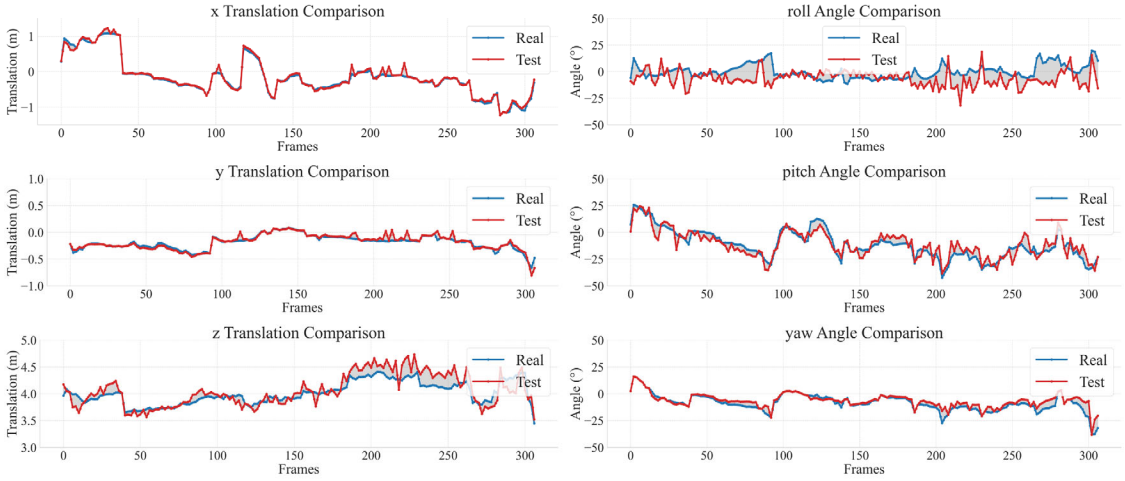

Figure 13. Comparison of tow navigation six-axis localization results with true trajectory. The left side shows translational comparisons in the *X*-, *Y*-, and *Z*-directions, and the right side shows rotational comparisons for roll, pitch, and yaw. The blue line represents the true values while the red line indicates the estimated results.

6.2. Autonomous Navigation Experiment Using Two AUVs

To further evaluate the localization accuracy and stability of the AUV6D model in multi-AUV cooperative operations, an autonomous navigation experiment was conducted using two AUVs from the "TS-MINI" series, specifically AUV12 and AUV14. These two vehicles were selected as the platforms for this experiment.

Experimental Setup: AUV12 and AUV14 navigated side by side in a 100 m-long test water area, each equipped with five cameras covering the front, up, down, left, and right directions. Figure 14 displays the overall setup. After deployment, both vehicles entered each other's visual field, submerged to a depth of 2.5 m, and proceeded under depth-controlled navigation. During this phase, the left-side camera of AUV14 and the right-side camera of AUV12 captured image data of each other, which were used as input for mutual pose estimation by the AUV6D model.

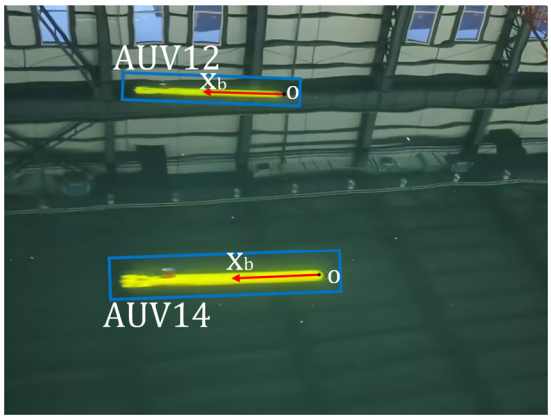

Figure 14. Experimental setup illustration. AUV12 and AUV14 during the coordinated navigation localization experiment, with each AUV's coordinate axis (X_b) and origin (O) marked as reference points.

Given that the cameras were mounted on the heads of the AUVs, they appeared off-center in each other's video images during side-by-side navigation. As the navigation progressed, AUV12 moved more slowly, gradually exiting the field of view of AUV14 while AUV14 remained within the visual range of AUV12. At the end of the navigation, both vehicles re-entered each other's view. Figure 15 shows the sequence of images captured by the two AUVs throughout the navigation.

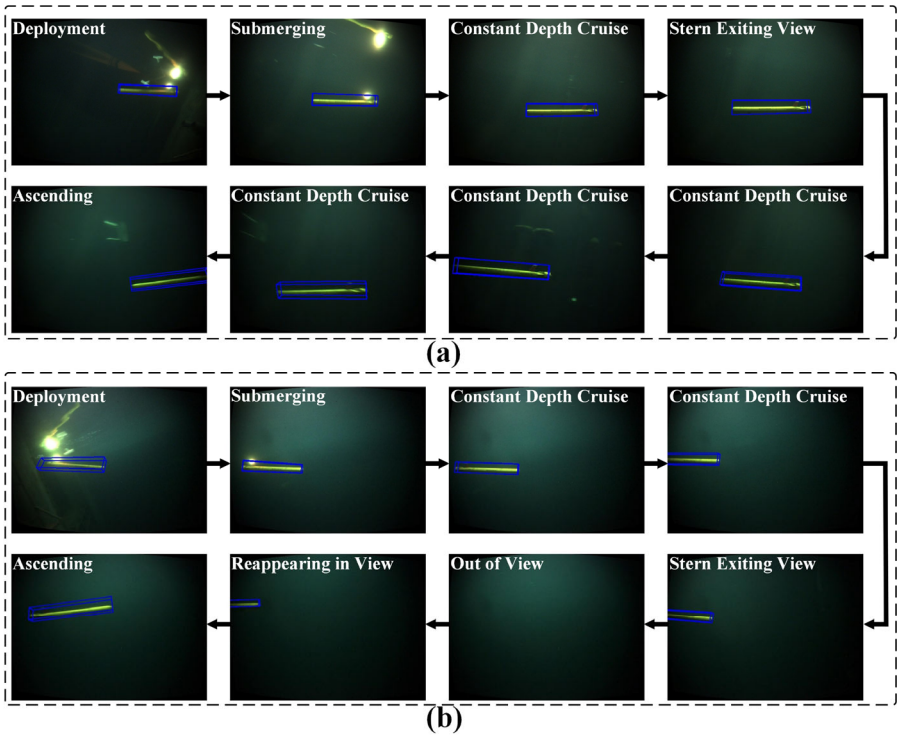

Figure 15. Image sequence captured during cooperative navigation of "TS-MINI" AUVs. (**a**) Sequence of AUV14captured by AUV12, from deployment, submergence, and depth-controlled navigation to the end of navigation. (**b**) Sequence of AUV12captured by AUV14, from deployment, submergence, and depth-controlled navigation to disappearing from view and reappearing.

To validate the reliability of the AUV6D model's inter-AUV pose estimation capability, the mutual 6D pose estimation results of the two AUVs were compared throughout the navigation process. This validation was achieved by ensuring the consistency of the 6D pose estimation results between the AUVs. Since each AUV's 6D pose estimation

values referenced their own coordinate systems, with the head as the origin (as shown in Figure 2), the relative pose data of one AUV with respect to the other were inversely related. Therefore, by reversing the six-degrees-of-freedom pose data of AUV14 relative to AUV12, the consistency and overlap of the data trends can be directly compared to reflect the reliability of mutual localization. A line chart of the mutual six-axis localization data was plotted in this manner, as shown in Figure 16.

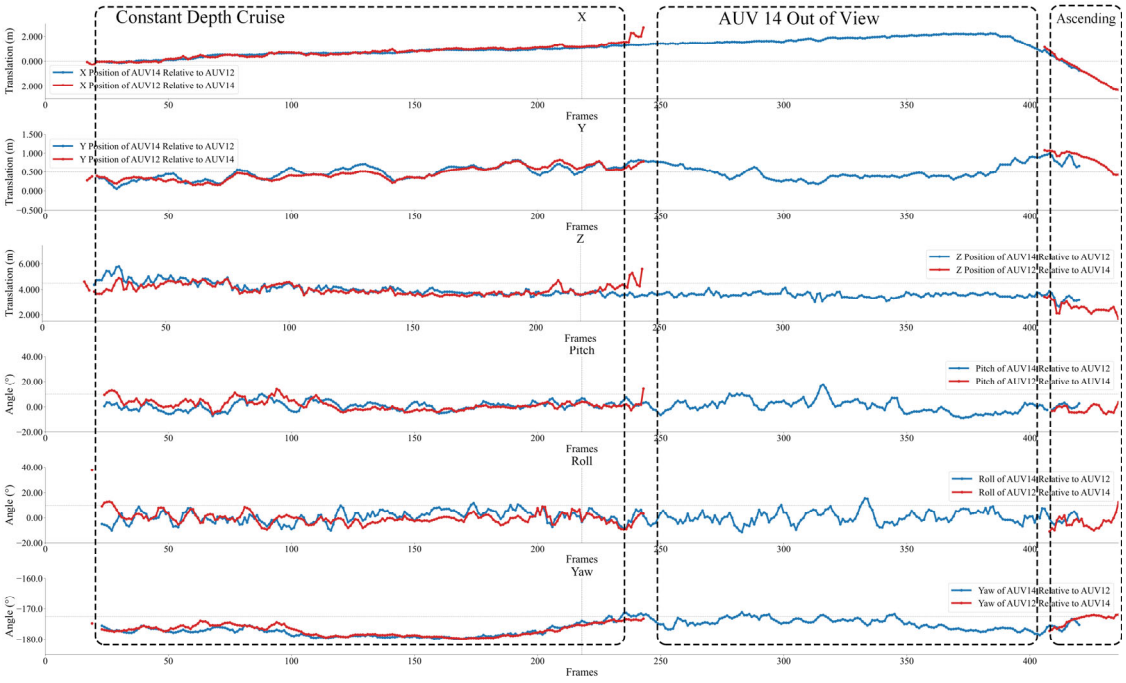

Figure 16. Comparative chart of relative pose estimation during cooperative navigation (with AUV14 data inverted). The chart illustrates the six-degrees-of-freedom pose changes relative to each other during cooperative navigation, including translational data in the *X*-, *Y*-, and *Z*-directions, and rotational data for roll, pitch, and yaw. The blue line represents the pose of AUV14 relative to AUV12, and the red line represents the pose of AUV12 relative to AUV14.

The results demonstrate that, during autonomous navigation, both AUVs were able to consistently detect and estimate each other's position and orientation across all six degrees of freedom: the *X*-, *Y*-, and *Z*-directions, roll, pitch, and yaw. The trajectory curves also clearly reflect the moment when AUV14 gradually moved out of AUV12's field of view, followed by its reappearance later in the navigation. A comparison of the mutual pose estimation results showed high consistency between the two AUVs, with the six-axis localization data exhibiting minimal deviation and remaining within acceptable error margins. This consistency validates the AUV6D model's ability to provide reliable and accurate inter-AUV localization information, further confirming its suitability for use in multi-AUV swarm operations.

7. Conclusions

This study introduced the AUV6D model to address the challenges of inter-AUV 6D pose estimation in dynamic underwater environments. To overcome the scarcity of real underwater data, a comprehensive dataset of simulated underwater images was generated, and Mask-CycleGAN was employed to transform these into realistic synthetic images. The Color Intermediate Domain Mapping strategy was introduced to improve the model's adaptability across diverse underwater conditions, while the Salient Keypoint Vector Voting Mechanism was developed to enhance the accuracy and robustness of pose estimation.

The experimental results demonstrated that the AUV6D model achieved millimeter-level localization precision and maintained pose estimation errors within five degrees. The

model adapted effectively to dynamic underwater environments, including variations in lighting, water bodies, and occlusion scenarios.

While the AUV6D model has shown promising results, several limitations should be acknowledged. The reliance on synthetic training data, despite the implementation of the Color Intermediate Domain Mapping strategy, resulted in reduced success rates when applied to environments with drastically different visual characteristics. Specifically, in experimental scenarios like Pool and Low Light (Lit), the model's performance decreased. Additionally, the model's behavior under extreme conditions, such as high-turbidity or deep-sea environments, remains untested and warrants further investigation.

Future Applications: To enhance its real-world applicability, future research could focus on developing lightweight architectures to reduce computational load and improve the real-time deployment of the AUV6D model in underwater robotic systems. The model could be extended to collaborative tasks in underwater robotics, such as target tracking, underwater docking, cooperative exploration, and autonomous operations. These applications would broaden the model's utility, enabling more complex multi-AUV operations in challenging environments.

Theoretical Advancements: On the theoretical front, further research should focus on enhancing the model's adaptability to more complex environmental conditions, enabling it to better adjust to varying scenarios as they arise. Additionally, investigating pose estimation for unknown objects presents a promising avenue to increase the model's versatility, enabling it to handle a wider range of operational scenarios beyond predefined objects.

In conclusion, this research provides a robust foundation for improving inter-AUV perception and localization. The methodologies and experimental validations presented in this study underscore the potential of the AUV6D model to enhance formation accuracy and collaboration in future AUV swarm deployments. Addressing the identified limitations and exploring the proposed future directions will be essential to fully realizing the potential of this approach in dynamic and challenging underwater environments.

Author Contributions: Conceptualization, Q.W. and Y.Y.; methodology, Q.W. and X.Z.; software, Q.W. and C.F.; validation, Q.W., Y.Y. and Z.W.; formal analysis, Y.Y. and X.Z.; investigation, Q.W.; resources, Y.Y., Q.Z. and Z.H.; data curation, Q.W.; writing—original draft preparation, Q.W.; writing—review and editing, Y.Y., X.Z. and Y.L.; visualization, Q.W.; supervision, Y.Y. and Z.H.; project administration, Y.Y. and Z.H.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDC 03060201.

Data Availability Statement: The data related to the reported results are currently not publicly available due to confidentiality restrictions.

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

References

1. Furfaro, T.C.; Alves, J. An Application of Distributed Long BaseLine—Node Ranging in an Underwater Network. In Proceedings of the Underwater Communications Networking (UComms), Sestri Levante, Italy, 3–5 September 2014.
2. Allotta, B.; Caiti, A.; Costanzi, R.; Di Corato, F.; Fenucci, D.; Monni, N.; Pugi, L.; Ridolfi, A. Cooperative navigation of AUVs via acoustic communication networking: Field experience with the Typhoon vehicles. *Auton. Robot.* **2016**, *40*, 1229–1244. [CrossRef]
3. Wang, Z.; Guan, X.; Liu, C.; Yang, S.; Xiang, X.; Chen, H. Acoustic communication and imaging sonar guided AUV docking: System and lake trials. *Control Eng. Pract.* **2023**, *136*, 105529. [CrossRef]
4. Fallon, M.F.; Papadopoulos, G.; Leonard, J.J. A Measurement Distribution Framework for Cooperative Navigation using Multiple AUVs. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Anchorage, AK, USA, 3–8 May 2010; pp. 4256–4263.
5. Kebkal, K.G.; Kabanov, A.A. Research on Feasibility of Low-Observable Acoustic Communication in AUV Group Navigation. *Gyroscopy Navig.* **2023**, *14*, 328–338. [CrossRef]
6. Zhuo, X.; Hu, T.; Wu, W.; Tang, L.; Qu, F.; Shen, X. Multi-AUV Collaborative Data Collection in Integrated Underwater Acoustic Communication and Detection Networks. In Proceedings of the IEEE Conference on Global Communications (IEEE GLOBECOM)—Intelligent Communications for Shared Prosperity, Kuala Lumpur, Malaysia, 4–8 December 2023; pp. 6771–6776.

7. Jiang, W.; Yang, X.; Tong, F.; Yang, Y.; Zhou, T. A Low-Complexity Underwater Acoustic Coherent Communication System for Small AUV. *Remote Sens.* **2022**, *14*, 3405. [CrossRef]
8. Tang, Z. Long Baseline Underwater Acoustic Location Technology. In *Encyclopedia of Ocean Engineering*; Cui, W., Fu, S., Hu, Z., Eds.; Springer: Singapore, 2020; pp. 1–6.
9. Zhao, W.; Qi, S.; Liu, R.; Zhang, G.; Liu, G. *A Review of Underwater Multi-source Positioning and Navigation Technology*; Springer: Singapore, 2023; pp. 5466–5479.
10. Tang, J.; Chen, Z.; Fu, B.; Lu, W.; Li, S.; Li, X.; Ji, X. ROV6D: 6D Pose Estimation Benchmark Dataset for Underwater Remotely Operated Vehicles. *IEEE Robot. Autom. Lett.* **2024**, *9*, 65–72. [CrossRef]
11. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 27–30 June 2016; pp. 779–788.
12. Jian, M.; Yang, N.; Tao, C.; Zhi, H.; Luo, H. Underwater object detection and datasets: A survey. *Intell. Mar. Technol. Syst.* **2024**, *2*, 9. [CrossRef]
13. Xu, S.; Zhang, M.; Song, W.; Mei, H.; He, Q.; Liotta, A. A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing* **2023**, *527*, 204–232. [CrossRef]
14. Zhou, J.; Pang, L.; Zhang, D.; Zhang, W. Underwater Image Enhancement Method via Multi-Interval Subhistogram Perspective Equalization. *IEEE J. Ocean. Eng.* **2023**, *48*, 474–488. [CrossRef]
15. Ruan, J.; Kong, X.; Huang, W.; Yang, W. Retiformer: Retinex-Based Enhancement In Transformer For Low-Light Image. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
16. Chen, X.; Liu, Y.; Wei, J.; Wan, Q.; Liu, S.; Cao, S.; Yin, X. Underwater image enhancement using CycleGAN. In Proceedings of the NCIT 2022; Proceedings of International Conference on Networks, Communications and Information Technology, Virtual, 5–6 November 2022; pp. 1–5.
17. Li, C.; Anwar, S.; Porikli, F. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognit.* **2020**, *98*, 107038. [CrossRef]
18. Feng, J.; Yao, Y.; Wang, H.; Jin, H. Multi-AUV Terminal Guidance Method Based On Underwater Visual Positioning. In Proceedings of the 2020 IEEE International Conference on Mechatronics and Automation (ICMA), Beijing, China, 13–16 October 2020; pp. 314–319.
19. Wei, Q.; Yang, Y.; Zhou, X.; Fan, C.; Zheng, Q.; Hu, Z. Localization Method for Underwater Robot Swarms Based on Enhanced Visual Markers. *Electronics* **2023**, *12*, 4882. [CrossRef]
20. Zhang, L.; Li, Y.; Pan, G.; Zhang, Y.; Li, S. Terminal Stage Guidance Method for Underwater Moving Rendezvous and Docking Based on Monocular Vision. In Proceedings of the OCEANS 2019, Marseille, France, 17–20 June 2019; pp. 1–6.
21. Di, Y.; Manhardt, F.; Wang, G.; Ji, X.; Navab, N.; Tombari, F. SO-Pose: Exploiting Self-Occlusion for Direct 6D Pose Estimation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 10–17 October 2021; pp. 12376–12385.
22. Hu, Y.; Fua, P.; Wang, W.; Salzmann, M. Single-Stage 6D Object Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, Seattle, WA, USA, 14–19 June 2020; pp. 2927–2936.
23. Wang, G.; Manhardt, F.; Tombari, F.; Ji, X. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Hefei, China, 20–25 June 2021; pp. 16606–16616.
24. Peng, S.; Zhou, X.; Liu, Y.; Lin, H.; Huang, Q.; Bao, H. PVNet: Pixel-Wise Voting Network for 6DoF Object Pose Estimation. *Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3212–3223. [CrossRef] [PubMed]
25. Song, C.; Song, J.; Huang, Q. HybridPose: 6D Object Pose Estimation under Hybrid Representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, Seattle, WA, USA, 14–19 June 2020; pp. 428–437.
26. Tekin, B.; Sinha, S.N.; Fua, P. Real-Time Seamless Single Shot 6D Object Pose Prediction. In Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 292–301.
27. Liu, P.; Zhang, Q.; Zhang, J.; Wang, F.; Cheng, J. MFPN-6D: Real-time One-stage Pose Estimation of Objects on RGB Images. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Xian, China, 30 May–5 June 2021; pp. 12939–12945.
28. Zhang, S.; Zhao, W.; Guan, Z.; Peng, X.; Peng, J. Keypoint-graph-driven learning framework for object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Electr Network, Nashville, TN, USA, 19–25 June 2021; pp. 1065–1073.
29. Zheng, Y.; Zheng, C.; Shen, J.; Liu, P.; Zhao, S. Keypoint-Guided Efficient Pose Estimation and Domain Adaptation for Micro Aerial Vehicles. *IEEE Trans. Robot.* **2024**, *40*, 2967–2983. [CrossRef]
30. Joshi, B.; Modasshir, M.; Manderson, T.; Damron, H.; Xanthidis, M.; Li, A.Q.; Rekleitis, I.; Dudek, G. DeepURL: Deep Pose Estimation Framework for Underwater Relative Localization. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 1777–1784.

31. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2242–2251.
32. Sun, B.; Jia, S.; Jiang, X.; Jia, F. Double U-Net CycleGAN for 3D MR to CT image synthesis. *Int. J. Comput. Assist. Radiol. Surg.* **2023**, *18*, 149–156. [CrossRef] [PubMed]
33. Gong, C.; Huang, Y.; Luo, M.; Cao, S.; Gong, X.; Ding, S.; Yuan, X.; Zheng, W.; Zhang, Y. Channel-wise attention enhanced and structural similarity constrained cycleGAN for effective synthetic CT generation from head and neck MRI images. *Radiat. Oncol.* **2024**, *19*, 37. [CrossRef] [PubMed]
34. Li, J.; Skinner, K.A.; Eustice, R.M.; Johnson-Roberson, M. WaterGAN: Unsupervised Generative Network to Enable Real-Time Color Correction of Monocular Underwater Images. *IEEE Robot. Autom. Lett.* **2018**, *3*, 387–394. [CrossRef]
35. Zhou, W.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef]
36. Fardo, F.A.; Conforto, V.H.; Oliveira, F.C.d.; Rodrigues, P.S.S. A Formal Evaluation of PSNR as Quality Measurement Parameter for Image Segmentation Algorithms. *arXiv* **2016**, arXiv:1605.07116.
37. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image Style Transfer Using Convolutional Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
38. Bynagari, N.B. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv* **2019**, arXiv:1706.08500. [CrossRef]