remote sensing

MDPI

Article

Reliable and Effective Stereo Matching for Underwater Scenes

Lvwei Zhu, Ying Gao *, Jiankai Zhang, Yongqing Li and Xueying Li

School of Data Science, Qingdao University of Science and Technology, Qingdao 266061, China; 4023112073@mails.qust.edu.cn (L.Z.); 4024181033@mails.qust.edu.cn (J.Z.); liyongqing@qust.edu.cn (Y.L.); lixueying@qust.edu.cn (X.L.)

* Correspondence: gaoying@qust.edu.cn

Abstract: Stereo matching plays a vital role in underwater environments, where accurate depth estimation is crucial for applications such as robotics and marine exploration. However, underwater imaging presents significant challenges, including noise, blurriness, and optical distortions that hinder effective stereo matching. This study develops two specialized stereo matching networks: UWNet and its lightweight counterpart, Fast-UWNet. UWNet utilizes self- and cross-attention mechanisms alongside an adaptive 1D-2D cross-search to enhance cost volume representation and refine disparity estimation through a cascaded update module, effectively addressing underwater imaging challenges. Due to the need for timely responses in underwater operations by robots and other devices, real-time processing speed is critical for task completion. Fast-UWNet addresses this challenge by prioritizing efficiency, eliminating the reliance on the time-consuming recurrent updates commonly used in traditional methods. Instead, it directly converts the cost volume into a set of disparity candidates and their associated confidence scores. Adaptive interpolation, guided by content and confidence information, refines the cost volume to produce the final accurate disparity. This streamlined approach achieves an impressive inference speed of 0.02 s per image. Comprehensive tests conducted in diverse underwater settings demonstrate the effectiveness of both networks, showcasing their ability to achieve reliable depth perception.

Keywords: underwater imaging; stereo matching; depth estimation

check for updates

Citation: Zhu, L.; Gao, Y.; Zhang, J.; Li, Y.; Li, X. Reliable and Effective Stereo Matching for Underwater Scenes. *Remote Sens.* **2024**, *16*, 4570. https://doi.org/10.3390/rs16234570

Academic Editor: Jaroslaw Tegowski

Received: 31 October 2024
Revised: 29 November 2024
Accepted: 3 December 2024
Published: 5 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, stereo matching has emerged as a pivotal technique in the field of computer vision, particularly for applications requiring 3D scene perception. This is especially true in underwater environments, where the demand for accurate depth estimation is driven by various fields, including robotics [1,2], marine exploration [3], and object recognition [4]. The ability to perceive three-dimensional structures underwater is crucial for navigating complex marine terrains, conducting environmental assessments, and identifying and tracking objects in real time. However, underwater imaging presents significant challenges that complicate the stereo matching process. Factors such as blurriness, noise, and visual distortions caused by the unique optical properties of water can severely degrade image quality [5], making it difficult to accurately identify matching points across stereo pairs. These challenges underscore the need for precise stereo matching techniques, as the accuracy of depth estimation directly influences the effectiveness of applications in underwater robotics and exploration. Thus, advancing stereo matching methodologies that can effectively mitigate these challenges is essential for enhancing the performance of underwater perception systems.

Recent advances in stereo matching networks have demonstrated promising results in various applications. However, many of these methods struggle to cope with the unique challenges presented by underwater environments, where domain-specific stereo matching approaches remain limited. For instance, ref. [6] proposes a belief-propagation-based stereo matching method that incorporates an energy function tailored to underwater scenarios.

By measuring the similarity between stereo image pairs, this approach aims to mitigate disparity map discontinuities. However, in real-world underwater scenes, it struggles to delineate object boundaries accurately due to imaging distortions, failing to produce clear disparity edges. Additionally, ref. [7] introduces style adaptation, semantic adaptation, and disparity range adaptation to incrementally tailor terrestrial depth estimation models for underwater domains. This approach integrates adaptation modules into existing stereo matching methods such as [8,9], improving the performance of state-of-the-art stereo matching models in underwater environments. Nevertheless, the method exhibits limitations in reflective and low-texture areas commonly found in real underwater scenes.

Despite continuous advancements in stereo matching technology, traditional methods often struggle when applied to underwater imagery due to the inherent challenges posed by these unique environments. Existing stereo matching algorithms are primarily designed for terrestrial applications, relying on assumptions such as uniform illumination, photometric consistency, sharp image features, and valid epipolar geometry. However, these assumptions often do not hold in underwater environments, resulting in inaccurate disparity maps and unreliable depth perception. In feature extraction, most stereo matching models for ground-based scenarios rely on deep convolutional networks to effectively capture multi-scale and contextual information [10–12]. However, they often overlook global information and the mutual relationship between image pairs, both of which are crucial for underwater scenes. Underwater images are profoundly impacted by water turbulence and suspended particles. Moreover, the imaging process in underwater environments is heavily influenced by water properties and light conditions, often leading to blurriness, noise, and even distortion, which obscure critical features. Traditional stereo matching methods struggle to extract accurate and robust feature information from such challenging images, impeding the matching process. To address these challenges, we introduce self-attention and cross-attention mechanisms to better leverage global information and the interrelationship between left and right images, laying a solid foundation for constructing a more accurate cost volume. Additionally, we utilize a cascaded iterative refinement process during the disparity update phase, enabling the model to better exploit the high-receptive-field global information available at lower resolutions [13]. Furthermore, most stereo matching methods for terrestrial scenes constrain the matching process of epipolar-rectified stereo image pairs to 1D [10–12,14]. However, this simplification is unsuitable for underwater environments. Due to the inherent characteristics of underwater imaging, stereo image pairs often exhibit distortions that epipolar rectification alone cannot address, making it impractical to limit matching points to a single line. Consequently, 2D search becomes crucial in underwater scenarios. To address this issue, we propose a 1D-2D cross-search strategy, aiming to mitigate the limitations and enhance matching accuracy in underwater environments.

In addition, many existing models rely on iterative refinement processes [10–12] that significantly increase computational burdens. In each iteration of the loop, the network performs numerous interpolation, aggregation, or convolution operations, updating the cost volume, weights, or other intermediate features through repeated pixel-wise calculations. These intermediate features are high-resolution inputs with multiple channels, making these operations highly time-consuming [15,16]. Additionally, each iteration requires data to be loaded from GPU memory, computations, and the storage of results back into memory. Frequent data transfers and cache refreshes further contribute to latency. As the number of updates increases, the cumulative time cost for each iteration becomes substantial, resulting in significant overall computational overhead. This makes them unsuitable for real-time applications, particularly on platforms with limited processing capabilities, which can be a critical drawback in resource-constrained underwater settings. For instance, many Autonomous Underwater Vehicles (AUVs) operate on resource-constrained embedded systems to perform depth estimation tasks [17]. This situation creates a pressing need for methodologies that strike a balance between accuracy, speed, and robustness, ensuring reliable depth perception in challenging conditions. Our work aims to address these gaps

by introducing a lightweight stereo matching network that eliminates the need for iterative cycles to obtain the final disparity. By optimizing the network architecture and computational process, we employ strategies such as cost propagation and confidence-guided refinement, which enhance performance while minimizing computational complexity.

To tackle the challenges in underwater imaging, such as blurriness and distortion caused by the underwater environment, we propose a robust stereo matching model named UWNet. This network is designed to meet the precision requirements of stereo matching in underwater scenes, enhancing disparity estimation accuracy despite these challenging conditions. Additionally, we propose a novel lightweight stereo matching network named Fast-UWNet that integrates innovative strategies to enhance both efficiency and accuracy. Our approach significantly reduces computational complexity while maintaining, or even improving, the precision of depth estimation in challenging underwater conditions. We validated the effectiveness of our method in three underwater datasets: the first one is a real underwater scene dataset that we built by ourselves in our previous work, the second one is a publicly available real underwater dataset [18], and the third dataset used in this study is the Scene Flow dataset [19], which we processed using a style transfer method to transform the image pairs into underwater-style images. Additionally, we applied image enhancement techniques to eliminate image artifacts, thereby facilitating more accurate disparity estimation. This processed dataset allows us to validate the effectiveness of our method across a wide range of underwater scenarios. The main contributions of this paper are summarized as follows:

- We propose a stereo matching network named UWNet tailored for underwater scenes, employing both self-attention and cross-attention mechanisms in tandem to capture accurate feature representations despite the high noise levels typical in underwater images. To address underwater image distortion, we utilize a 1D-2D cross-search to adaptively expand the expressive range of the cost volume. Additionally, a cascaded recurrent update module is applied to progressively refine the disparity from lower resolutions, enabling the final precise error estimation.
- Building on UWNet, we further propose Fast-UWNet, a lightweight stereo matching network with an inference speed of only 0.02 s. Fast-UWNet directly generates low-resolution disparity candidates and confidence scores from the cost volume, minimizing the need for iterative updates. This design leverages confidence-driven interpolation, adaptively refining both content and confidence information to incrementally correct cost volume errors, ultimately achieving precise disparity estimation.
- We validate the effectiveness of UWNet and Fast-UWNet across diverse underwater environments, with experiments demonstrating the superiority of our approach.

2. Methods

In addition to utilizing publicly available underwater stereo matching datasets, we employ style transfer techniques to augment the quantity of underwater-style datasets. This expansion enriches the dataset available for model training and testing. Furthermore, we apply image enhancement methods to the transformed images to reduce artifacts, facilitating more accurate disparity estimation. Most importantly, we propose a novel network named UWNet that effectively addresses the challenges of underwater stereo matching, achieving precise disparity estimation. Building upon this foundational framework, we further introduce a lightweight network, Fast-UWNet, optimized for efficient underwater stereo matching.

2.1. Style Transfer

Due to the limited availability of real-world underwater stereo matching datasets, we use style transfer to artificially expand the dataset. While style transfer does not directly increase the number of original samples, it generates additional variations of the same images by adapting their visual appearance to that of underwater scenes. This process allows the model to learn from a broader range of image characteristics, improving its

ability to generalize to real underwater environments. Additionally, to eliminate artifacts introduced during the style transfer process and ensure high-quality images for the stereo matching network, we apply non-local means filtering as a post-processing step [20–22].

Style transfer works by separating the content and style of an image, where the content represents the structure of the image and the style refers to its visual characteristics, such as the unique lighting and texture properties found in underwater environments [23–26]. The transformed images maintain the original structure while adopting the desired underwater style, as shown in Figure 1. To achieve this, we leverage Adaptive Instance Normalization (AdaIN) [27–29], a technique designed for effective style transfer. AdaIN begins by utilizing VGG-19 [30], a pre-trained neural network, to extract high-level feature maps from both the content image (x) and the style image (y). Let F_x and F_y denote the extracted feature maps of the content and style images, respectively. It then aligns the channel-wise mean $\mu(F_x)$ and variance $\sigma(F_x)$ of the content image to match those of the style image($\mu(F_y)$ and $\sigma(F_y)$), effectively blending the style into the content. This process enables seamless style transformation while preserving the structural integrity of the original image. Additionally, we determined the optimal content and style weightings specifically tailored for underwater environments during the style transfer training process. These weightings, adjusted through the loss function, balance the influence of the image's content and style in the transformed output, ensuring accurate and realistic domain adaptation.



(**a**) Content (**b**) Style

(**c**) Content+Style (**d**) Filter-treated

Figure 1. Visualization of the style-transferred images: (**a**) the original input image, (**b**) the input style image, (**c**) the style-transferred image after initial AdaIN processing, and (**d**) the image after non-local mean filtering.

For an input tensor of size B,C,H,W, the dimensions are interpreted as follows: B represents the batch size, and C denotes the number of channels, while H and W correspond to the height and width of the image, respectively. ϵ denotes the smoothing factor, which is used to avoid numerical instability due to a zero denominator or variance when calculating the standard deviation, as shown in following equation:

$$\mu_{bc}(F) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} F_{bchw} \tag{1}$$

$$\sigma_{bc}(F) = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(F_{bchw} - \mu_{bc}(F))^2 + \epsilon} \tag{2}$$

$$\text{AdaIN}(F_x, F_y) = \sigma(F_y)\left(\frac{F_x - \mu(F_x)}{\sigma(F_x)}\right) + \mu(F_y) \tag{3}$$

The loss function used during the style transfer process is expressed as follows. Here, F_t denotes the feature map of the transformed image. Through parameter tuning and extensive experimental testing, these values, $\alpha = 2.0$ and $\beta = 8.0$, were found to be particularly effective for underwater-style image transfer.

$$\mathcal{L}_{\text{content}} = \|F_x - F_t\|_2^2 \tag{4}$$

$$\mathcal{L}_{\text{style}} = \|\mu(F_x) - \mu(F_y)\|_2^2 + \|\sigma(F_x) - \sigma(F_y)\|_2^2 \tag{5}$$

$$\mathcal{L} = \alpha\mathcal{L}_{\text{content}} + \beta\mathcal{L}_{\text{style}} \tag{6}$$

We incorporated a non-local mean filter to mitigate the unwanted noise introduced by the initial style transfer process. The core principle of non-local mean filtering lies in reducing noise by identifying similar pixel blocks across the entire image. For each pixel, the new value is computed as a weighted average of surrounding regions that exhibit similarity, with the weights determined based on how closely these regions resemble the target pixel. This approach ensures more precise noise reduction by leveraging global image information.

Given an image I, each pixel $I(p)$ with its new denoised value $I'(p)$ can be calculated by the following equation:

$$I'(p) = \frac{1}{Z(p)}\sum_{q\in\Omega} I(q)\cdot w(p,q) \tag{7}$$

$Z(p)$ is the normalization factor ensuring that the weights sum to 1. $q \in \Omega$ indicates the neighborhood of pixels considered around p (here, we are using a 7×7 window). $w(p,q)$ is the weight that determines the contribution of the pixel $I(q)$ to the computation of $I'(p)$. Higher weight indicates greater similarity between $I(p)$ and $I(q)$.

$$Z(p) = \sum_{q\in\Omega} w(p,q) \tag{8}$$

$$w(p,q) = e^{-\frac{\|I(p)-I(q)\|^2}{h^2}} \tag{9}$$

$\|I(p) - I(q)\|$ is the distance between two blocks of pixels, and h is a parameter that controls the similarity measure; here, we set it to 7.

2.2. UWNet

The overview of the UWNet is shown in Figure 2. UWNet primarily consists of three modules: feature extraction, cost volume computation, and a cascaded recurrent network. The feature extraction module is designed to capture multi-scale and contextual features from the input stereo image pair. It employs both self-attention and cross-attention mechanisms, which work together to produce more accurate feature representations. During cost volume computation, we introduce a 1D-2D cross-search to adaptively enhance the expressive capacity of cost information. Finally, a cascaded recurrent network iteratively refines the disparity to obtain the final precise disparity estimation.

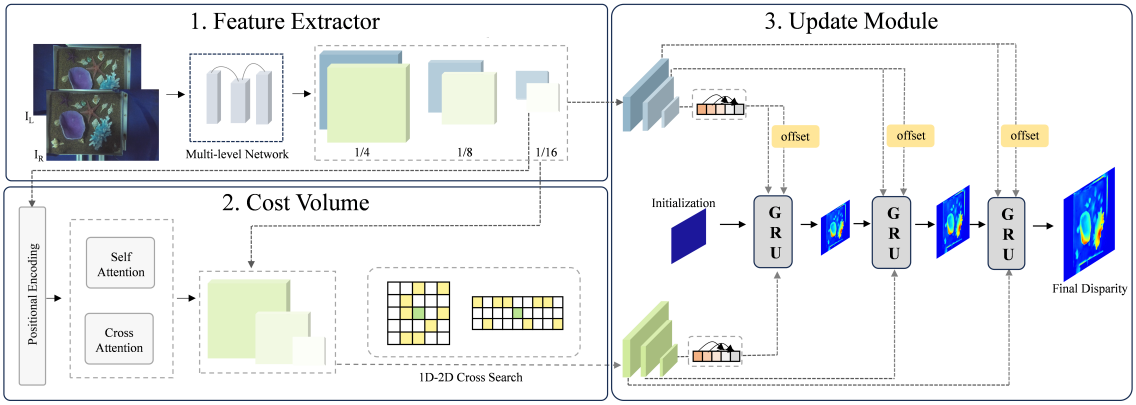


Figure 2. The overview of UWNet. The feature extractor processes the input stereo image pair to obtain feature information and contextual features. The lowest-resolution features are refined through the coordination of self-attention and cross-attention mechanisms to achieve more accurate feature representations. These processed features are then used to compute the cost volume through a 1D-2D cross-search strategy. Finally, the contextual features and the cost volume are fed into a cascaded iterative refinement module to generate the final disparity map.

2.2.1. Feature Extractor

Following [11], the feature extractor uses the MobileNetV2, pretrained on ImageNet [31]. It first scales down the input feature map $(f_{l,r})$ to $1/32$ of its original size and then employs upsampling blocks with skip connections to restore the features to a $1/4$ scale. This process results in multi-scale features $\{f_{l,i}, (f_{r,i}) \in R^{C_i \times \frac{H}{i} \times \frac{W}{i}}\}$, where $i = 4, 8, 16, 32$ and C_i denotes the number of feature channels. Additionally, following [10,11], the context extractor, composed of residual blocks and downsampling layers, generates context features at $1/4$, $1/8$, and $1/16$ scales.

Self-Attention and Cross-Attention. Inspired by [32], self-attention and cross-attention are employed to obtain a more accurate representation of features in underwater images. First, positional encoding is applied to the feature maps at $1/16$ resolution to enhance their spatial dependencies. Subsequently, through the collaborative use of self-attention and cross-attention mechanisms, global contextual information is aggregated either within individual feature maps or across the left and right feature maps. This approach effectively mitigates the impact of blurring and distortion in underwater images, resulting in more accurate and comprehensive feature representations.

2.2.2. Cost Volume

1D-2D cross-search. The cost volume represents the matching cost between corresponding points in the left and right images under a given disparity assumption. Due to the presence of suspended particles and reflective properties in water, underwater images often suffer from blurriness and distortion. As a result, a traditional 1D search to establish a cost volume between corresponding points is not sufficiently effective for underwater images. Inspired by [13], we introduce a 1D-2D cross-search method, as shown in Figure 3, which adaptively broadens the index range when constructing the cost volume, thereby minimizing the impact of distortions and improving the accuracy of stereo matching in underwater.

The cost is expressed using the product as shown in Equation (10), where $x' = x + dx + offset$ and $y' = y + dy + offset$. To ensure that the model is not disturbed by local noise, a sufficiently large search step is required without significantly increasing the computational burden. Therefore, in the 1D search, we set the search step range to $[-4,4]$. Accordingly, for the 2D search, we use a sliding window of size 3×3, where both search modes examine a total of 9 neighboring pixels, including the center pixel. The search is further adjusted with an offset as defined in Equation (11), aiming to minimize the adverse effects of underwater image distortions. We introduce a control parameter, determined by

the current iteration count, to regulate the search pattern at each iteration. This allows for the implementation of a cross-search strategy, combining both 1D and 2D searches.

$$Cost_{<x,y>} = f_L(x,y) * f_R(x',y') \tag{10}$$

$$offset = S * (sig(f_L) - 0.5) * 2 \tag{11}$$

In this case, S represents the size of the search step, and sig denotes the sigmoid function.

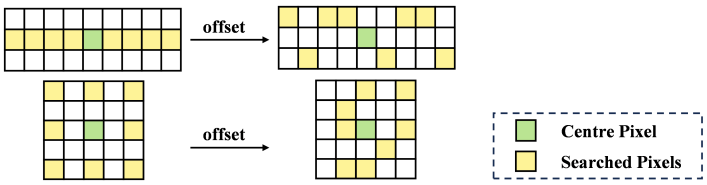


Figure 3. 1D-2D cross-search.

To reduce the number of parameters and save computational resources compared to calculating the cost volume separately at each scale, the cost volume is first established at a 1/4 scale, and then downsampling is performed to generate the cost volumes for the 1/8 and 1/16 scales.

2.2.3. Cascaded Recurrent Network

Low-resolution feature maps, which benefit from larger receptive fields and rich semantic information, tend to be more robust in these conditions but may lose finer structural details. To balance robustness and the preservation of high-resolution input details, we employ a cascaded recurrent refinement approach for updating disparity [13]. Unlike the recurrent iterative network in RAFT-Stereo [10], our method first computes disparity at the lowest resolution using the constructed cost pyramid. This initial estimate is then refined at each subsequent scale by combining lower resolution receptive field information with higher-dimensional features, ultimately producing the final disparity at the original resolution.

2.3. Fast-UWNet

Based on the UWNet, we constructed a lightweight stereo matching network, named Fast-UWNet. the overview of Fast-UWNet is shown in Figure 4. This method saves computing resources and enables real-time stereo matching.

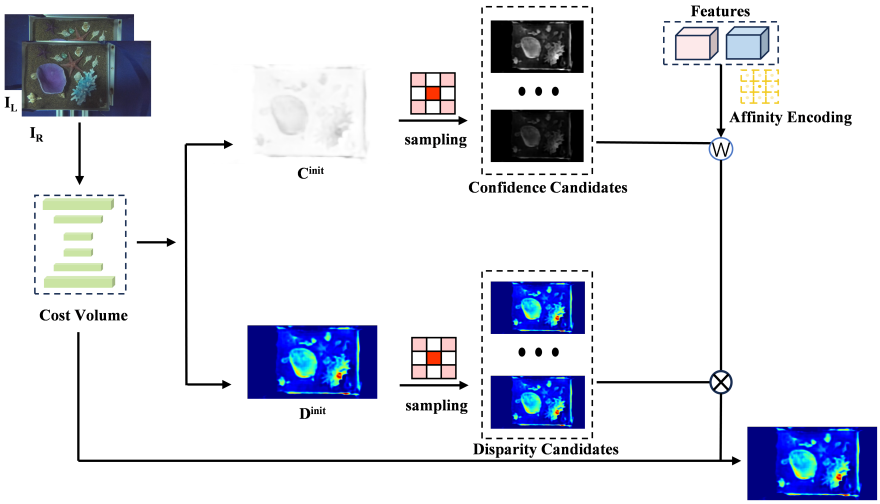


Figure 4. Overview of our Fast-UWNet. The core ideas are cost propagation and the affinity constraint. Disparity candidates and their confidence are generated based on cost propagation, and the affinity constraint effectively filters these candidates for improved accuracy.

2.3.1. Cost Propagation

Unlike methods that rely on multiple iterations of refinement through deep convolutional networks to produce the final disparity [10,12–14], our approach optimizes the model structure and computation process. The core idea is to generate a set of disparity candidates and their corresponding confidences based on cost propagation. Accurate interpolation is then performed adaptively using content information and confidence, gradually correcting cost volume errors to infer the final precise disparity estimation.

For the initial cost volume, we utilize the 1/8 resolution cost volume described in Section 2.2.2, which is then upsampled to 1/4 resolution for further use. Attention weights are generated using a softmax function, based on which a coarse disparity estimation and its confidence are derived. Through ablation experiments, to balance accuracy and efficiency, we employ a radial sampling strategy to select five disparity candidates and compute their respective confidences. As shown in Figure 5, next, affinity-constrained feature information and confidence are utilized as content information to adaptively perform precise interpolation on the cost volume. Using the updated attention weights, the optimal disparity estimation is eventually determined.

$$D^k = disp \times W^k \tag{12}$$

$$W^k = \text{SoftMax}(\text{Conv3D}(Cost^k_{<x,y>})) \tag{13}$$

$$Confidence^k = W^k \times (D^{init} - D^k)^2 \tag{14}$$

In the above equation, $disp$ is a sequence of disparity hypotheses ranging from 0 to $max_disp([0, 1, 2..., max_disp])$, used to map $W^k(k = 0, 1, 2, 3, 4)$ into D^k , which are then utilized to calculate the $confidence^k$ by comparing them with the D^{k-1}. W^k is generated from the cost volume using 3D convolution and a $SoftMax$ operation, ensuring normalized probabilities.

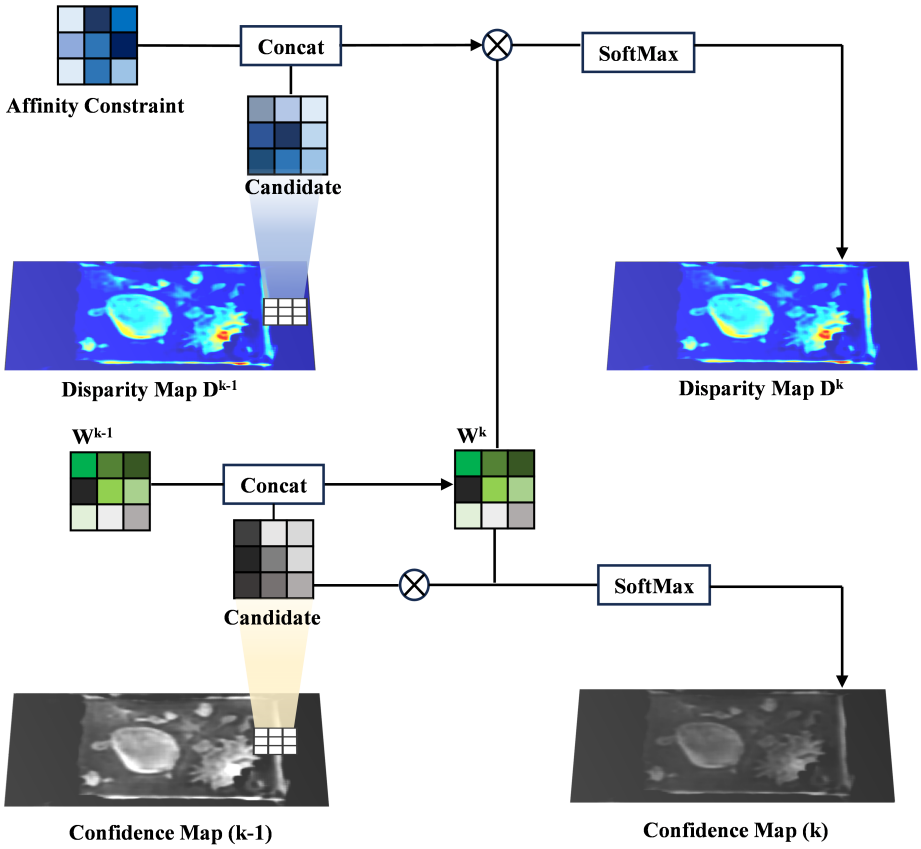


Figure 5. The k-th propagation process.

The result of the confidence diagram visualization is shown in Figure 6. This disparity candidate strategy compensates for mismatches caused by the blurring characteristics of underwater images and the presence of local reflective regions. Compared to a one-off assumption, the provision of candidate disparities prevents cumulative errors [33], continuously refining the attention weights to create more potential for accurate final disparity estimation.

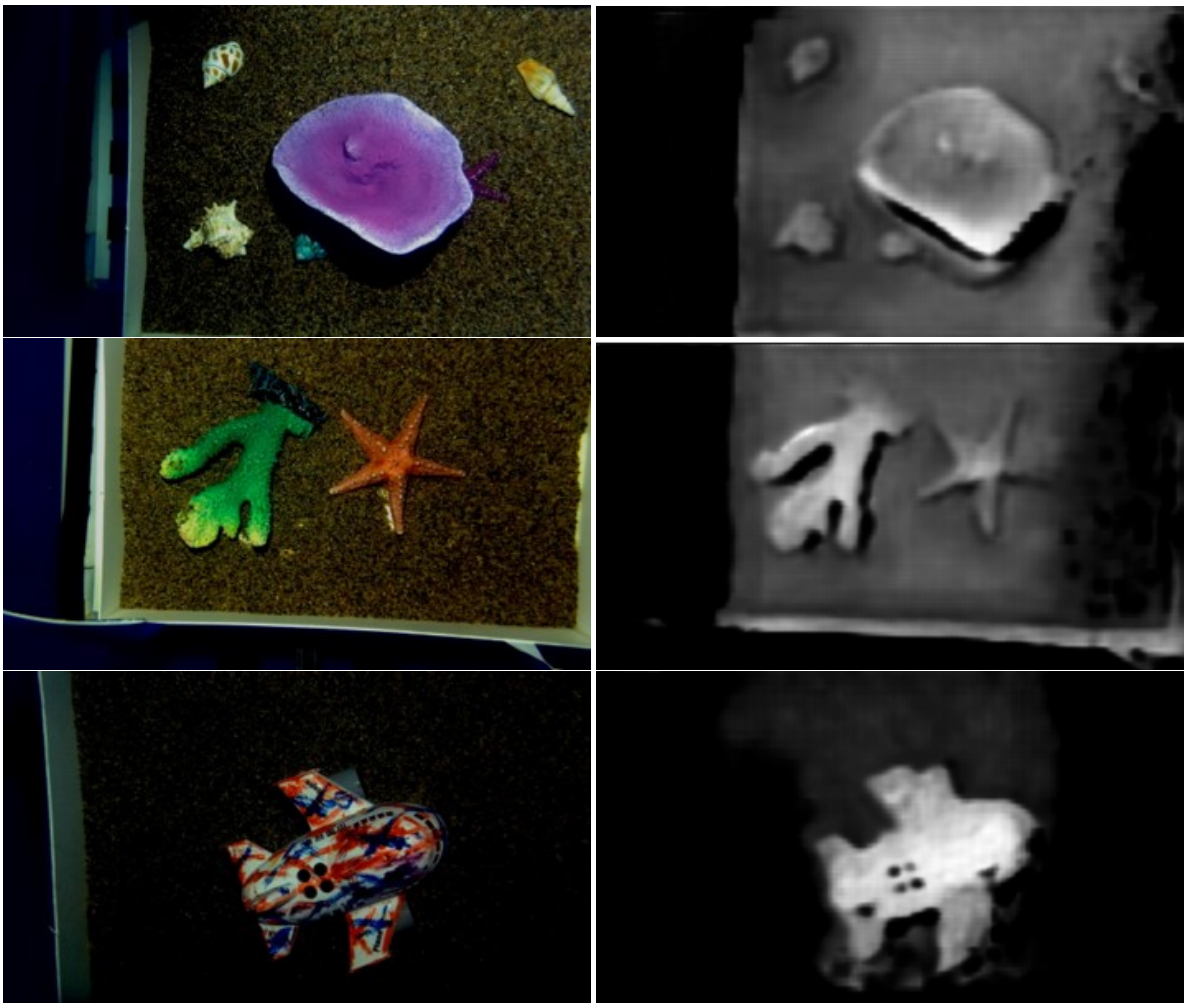
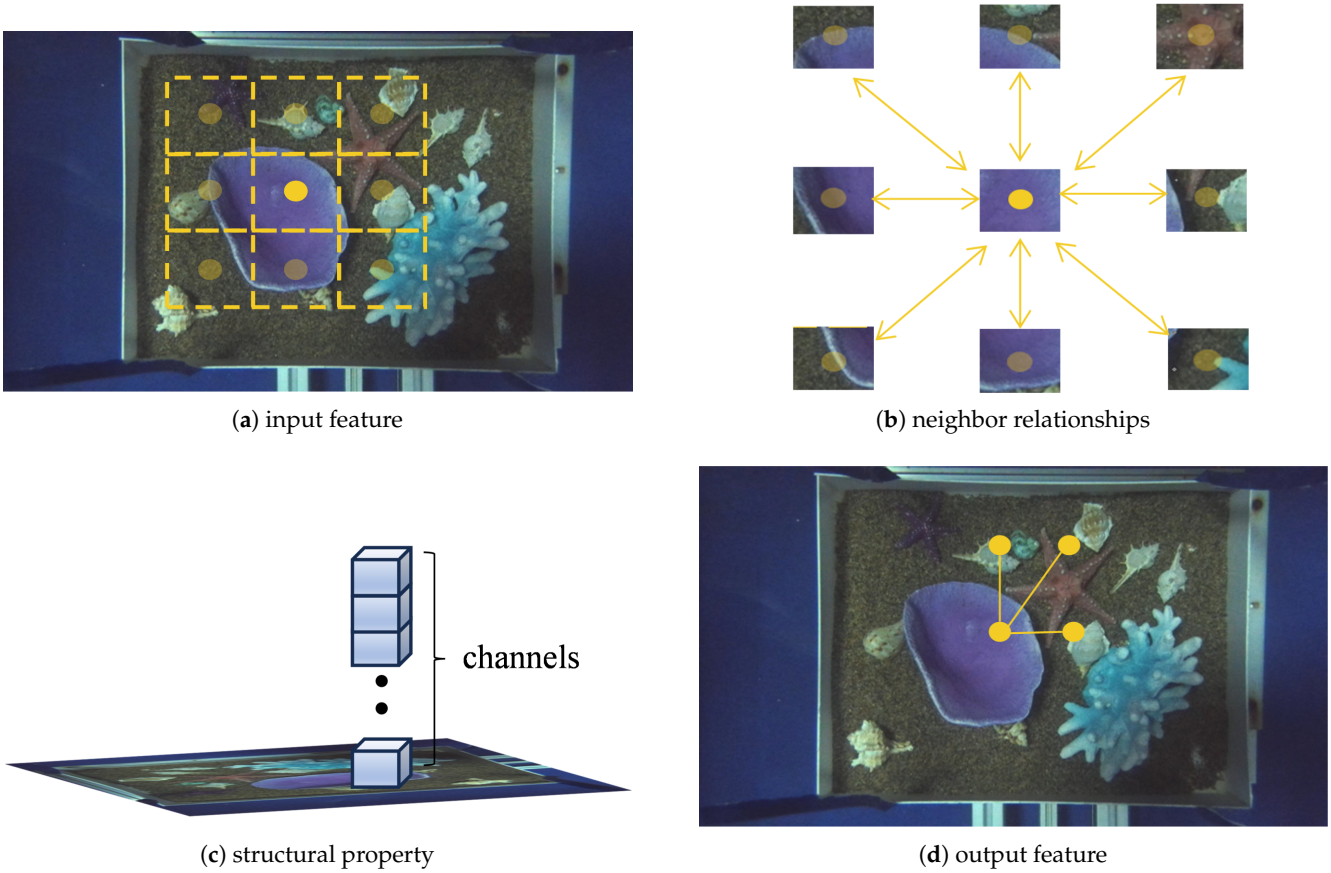


Figure 6. Visualization of confidence maps.

2.3.2. Affinity Constraint

Guided by the affinity constraint, as shown in Figure 7, the adjustment of attention weights limits disparity candidates to surrounding pixels that share an affinity with the central pixel, thus avoiding weight deviation caused by unrelated errors. The affinity relationship is achieved by applying structural constraints to the feature information, enforcing relationships among neighboring pixels [34,35]. This is accomplished by calculating the similarity between neighboring information and the central pixel. By setting a threshold, pixels with sufficient similarity are identified as affinity pixels of the central pixel.



(**a**) input feature (**b**) neighbor relationships

(**c**) structural property (**d**) output feature

Figure 7. The operational paradigm of the affinity constraint. Each yellow circle represents a pixel point in the feature map, and the relationship with neighboring pixels is calculated from the central pixel point outwards.

As shown in the following equation, where f_L^{aff} denotes the image feature information processed by the affinity constraint, *Norm* denotes the regularization normalization operation, and *Conv* denotes the convolution operation, which weights and sums the considered domain feature *neighbour* with the current feature f_L, reflecting the influence of the domain on the current point feature. *cut* denotes the threshold value, which is set to 0 to filter out irrelevant pixels.

$$f_L^{\text{aff}} = Norm\left(Conv\left(\sum(neighbor \times f_L)\right)\right) \tag{15}$$

$$f_L^{\text{aff}}[f_L^{\text{aff}} < cut] = cut \tag{16}$$

The high-likelihood weights from the previous level act as a filter to refine and derive new weights.

$$W^{k+1} = \text{Concat}\{Cost^k, f_L^{\text{aff}}, W^k\} \tag{17}$$

2.4. Loss Function

Based on the iterative output disparity-sequences: $\{d_1, ..., d_N\}$ and the initial disparity loss $\mathcal{L}_0$, the weights are defined to grow exponentially [10], where $\gamma = 0.9$, and the loss is defined according to the ground-truth disparity d_{gt}:

$$\mathcal{L}_0 = Smooth_{L1}(d_0 - d_{gt}) \tag{18}$$

d_0 is calculated from the initial cost volume *corr*:

$$d_0 = SoftMax(Conv(Corr)) \tag{19}$$

$$\mathcal{L} = \mathcal{L}_0 + \sum_{i=1}^{N} \gamma^{N-i} ||d_i - d_{gt}||_1 \quad (20)$$

3. Experiment

3.1. Datasets

Underwater Stereo Dataset. In this paper, we evaluate the applicability of our approach to underwater scenarios using a self-built dataset from previous work. This dataset simulates a real underwater scene with the camera positioned from the surface to the bottom, resulting in slight disparity variations. These conditions challenge the model's convergence and ability to refine details. Following cross-validation principles, we divided 49 image sets into 45 training pairs and four testing pairs. The images were captured using a ZED stereo camera enclosed in a waterproof casing, which was positioned at a depth of 1 m underwater. The camera recorded at a resolution of 1080 × 1920 pixels and operated at a frame rate of 60 fps.

SQUID [18]. The dataset includes 57 stereo pairs from four distinct sites in Israel, two of which are located in the Red Sea (representing tropical water conditions) and the other two in the Mediterranean Sea (representing temperate water conditions). We selected a subset of images from each location to form the test set for our experiments, the dataset was divided into 53 training sets and four test sets. This dataset presents challenges due to large disparity variations across extensive textureless regions. It is worth noting that the disparity values in the SQUID dataset are computed ground truth rather than direct in situ measurements, which may introduce potential biases.

Scene Flow [19]. Scene Flow is a synthetic stereo dataset collection that provides 35,454 training image pairs and 4370 testing image pairs, with a resolution of 540 × 960. The dataset provides dense disparity maps as ground truth. For our experiments, we utilized the Cleanpass subset of the Scene Flow dataset. We selected 500 samples from the complete dataset, which were transformed via style transfer to create our experimental dataset. In line with cross-validation principles, the dataset was divided into 450 samples for training and 50 samples for testing.

3.2. Implementation Details

We implemented our model using the PyTorch framework 1.13.0 and trained it with the AdamW optimizer. During the training process, we randomly cropped the images to a size of 320 × 736 (images were restored to full resolution during the evaluation process). In our UWNet experiments, we employed a one-cycle learning rate schedule, beginning with an initial learning rate of 0.0002, and implemented 20 update iterations throughout the training. In our Fast-UWNet experiments, we also employed a one-cycle learning rate schedule, beginning with an initial learning rate of 0.001. During all experiments, we used NVIDIA A10 GPUs.

3.3. Evaluation Metrics

In our experiments, we utilized End-Point Error (EPE) and D (D1, D3, D5, and D7) as evaluation metrics.

EPE measures the average pixel-wise discrepancy between the predicted disparity map and the ground truth. It provides a quantitative assessment of the accuracy of disparity estimation, where lower EPE values indicate better performance.

D metrics, on the other hand, assess the percentage of pixels for which the estimated disparity deviates from the ground truth by more than a certain threshold. These metrics (such as D1, D3, D5, and D7) provide a more comprehensive evaluation of the disparity estimation quality, capturing both small and larger errors.

3.4. Ablation Study

To validate and comprehensively understand the architecture of our model, we conducted a series of ablation experiments. All hyperparameters were kept consistent with the

training phase on the previously self-established underwater dataset. For UWNet in the ablation studies, we performed 20 iterations of updating at inference. The baseline used in these experiments utilized mobilenetv2_100 as the backbone for IGEV-Stereo [11]. For Fast-UWNet in the ablation studies, we trained the model with the key modules removed with the same steps as the full model. The results are shown in Table 1.

Table 1. Ablation study on UWNet.

Model	Attention	Cross 1D-2D Search	Cascaded	EPE (px)	>3px (%)
baseline				6.74	45.17
+attention	✓			6.20	40.32
+attention+cross	✓	✓		5.64	35.41
+cascaded			✓	6.35	40.81
UWNet	✓	✓	✓	4.34	31.57

Non-Local Mean Filter. We trained the network using both the unfiltered and filtered images as inputs, respectively. The results obtained from these two variations are shown in Table 2, highlighting the impact of filtering on the model's performance and the quality of disparity estimation.

Table 2. The ablation study on non-local mean filtering to evaluate its impact on the network's performance.

Input	EPE (px)	>1px (%)
unfiltered	2.39	35.83
Filter-treated	2.36	35.71

Effectiveness of Attention Mechanism. With the collaborative effect of self-attention and cross-attention mechanisms, the EPE decreased by 8.0%, and the D3 error was reduced by 10.7%. These attention mechanisms provided more reliable extraction of image features, effectively reducing errors caused by reflective regions and textureless areas, leading to improved accuracy in disparity estimation.

Effectiveness of 1D-2D Cross-search. After incorporating the 1D-2D cross-search, the EPE was reduced by 9.0%, and the D3 error decreased by 12.2%. This indicates that the proposed approach adaptively broadens the cost volume representation, effectively eliminating local ambiguities present in underwater images.

Effectiveness of Cascaded Recurrent Network. By replacing the iterative update module with the cascade recurrent update, the EPE decreased by 5.8% and the D3 error was reduced by 9.7%. This demonstrates that the cascade recurrent update performs better under challenging conditions where the image quality of underwater scenes is suboptimal.

Effectiveness of Disparity Candidate. Incorporating disparity candidates generated based on surrounding affinity pixels helps mitigate error accumulation that arises from producing a single disparity estimate. As shown in Table 3, the EPE decreased significantly by 18.1%, demonstrating the effectiveness of the module.

Table 3. Ablation study on Fast-UWNet.

Model	Candidate	Affinity	EPE (px)	>3px (%)	Time (ms)	Params (M)
baseline			7.74	58.65	49	3.20
+Candidate	✓		6.34	56.20	42	3.28
+Affinity		✓	6.98	57.85	36	3.22
Fast-UWNet	✓	✓	6.51	54.12	22	3.30

As shown in Table 4, to determine the optimal number of disparity candidates to use during training, we conducted ablation experiments by adjusting the number of selected

candidates on underwater dataset. When fewer than five candidates were selected, the EPE was reduced by approximately 15% for each increment. However, when more than five candidates were selected, the EPE reduction slowed significantly, dropping by only about 3% per additional candidate. Based on these observations, selecting five disparity candidates was identified as the optimal balance between accuracy and efficiency.

Table 4. Ablation study on the number of disparity candidates.

Candidates	2	3	4	5	6	7
EPE	8.72	7.34	6.91	**6.51**	6.22	6.14
Runtime (ms)	17	19	21	**22**	24	26

Effectiveness of Affinity Constraint. As shown in Table 3, the inclusion of affinity constraints from neighboring pixels in the weight adjustment process reduced EPE by 9.8%. This result demonstrates that the module effectively imposes structural constraints, supporting more accurate disparity estimation.

3.5. Comparisons with State of the Art

To evaluate our proposed stereo matching model for underwater scenes, we conducted comparisons against state-of-the-art methods on three datasets: our own real-world underwater dataset, the SQUID dataset, and an underwater-transferred version of the Scene Flow dataset. We used multiple evaluation metrics, including End Point Error (EPE), D3, D5, and D7, to assess each method's performance comprehensively.

As shown in Table 5, our UWNet outperformed RAFT-Stereo [10], DLNR [12], and IGEV-Stereo [11] across all four metrics (EPE, D3, D5, and D7). Key observations from different scenes highlight the advantages of our method. As shown in Figure 8, in the first scene, UWNet demonstrated greater sensitivity to subtle disparity variations, generating results closer to the ground truth and handling edge details more precisely than competing methods. In the second scene, for small objects, such as a starfish, our approach was able to produce the most complete and accurate disparity representation. In the third scene, in challenging shadowed areas, like the ear of a bear figure, UWNet successfully restored accurate disparities, even under low-contrast conditions.

Table 5. Quantitative evaluation on underwater dataset.

Method	EPE (px)	>3px (%)	>5px (%)	>7px (%)	Time (ms)	Params (M)
RAFT-Stereo [10]	6.86	46.77	34.59	26.29	132	11.23
DLNR [12]	6.03	42.58	30.41	22.83	297	57.37
IGEV-Stereo [11]	6.74	45.17	33.46	26.13	180	12.62
Fast-ACVNet [33]	7.74	58.64	54.21	50.32	49	5.22
UWNet	**4.34**	**31.57**	**21.88**	**16.14**	270	12.05
Fast-UWNet	**6.51**	**54.81**	**47.61**	**41.20**	**22**	3.30

The SQUID dataset contains a large number of suspended particles and large untextured areas, which makes it difficult for traditional models to perform. As shown in Table 6, UWNet consistently outperformed RAFT-Stereo [10], DLNR [12], and IGEV-Stereo [11] across all evaluation metrics. As shown in Figure 9, our method showed particularly robust performance in challenging textureless regions, where traditional models often struggle to generate reliable disparity estimates. Despite the difficulties that lightweight models face with this dataset, our Fast-UWNet still achieved better metric scores than Fast-ACVNet, although both models faced some limitations in visualizing fine disparity details.

Table 6. Quantitative evaluation on SQUID dataset.

Method	EPE (px)	>3px (%)	>5px (%)	>7px (%)	Time (ms)	Params (M)
RAFT-Stereo [10]	5.17	57.34	42.34	27.25	132	11.23
DLNR [12]	4.95	55.71	38.12	24.08	297	57.37
IGEV-Stereo [11]	4.80	53.50	35.13	20.01	180	12.62
Fast-ACVNet [33]	4.79	60.20	39.34	25.22	49	5.22
UWNet	**3.74**	**44.42**	**26.23**	**13.98**	270	12.05
Fast-UWNet	**4.47**	**58.53**	**32.82**	**23.16**	**22**	3.30

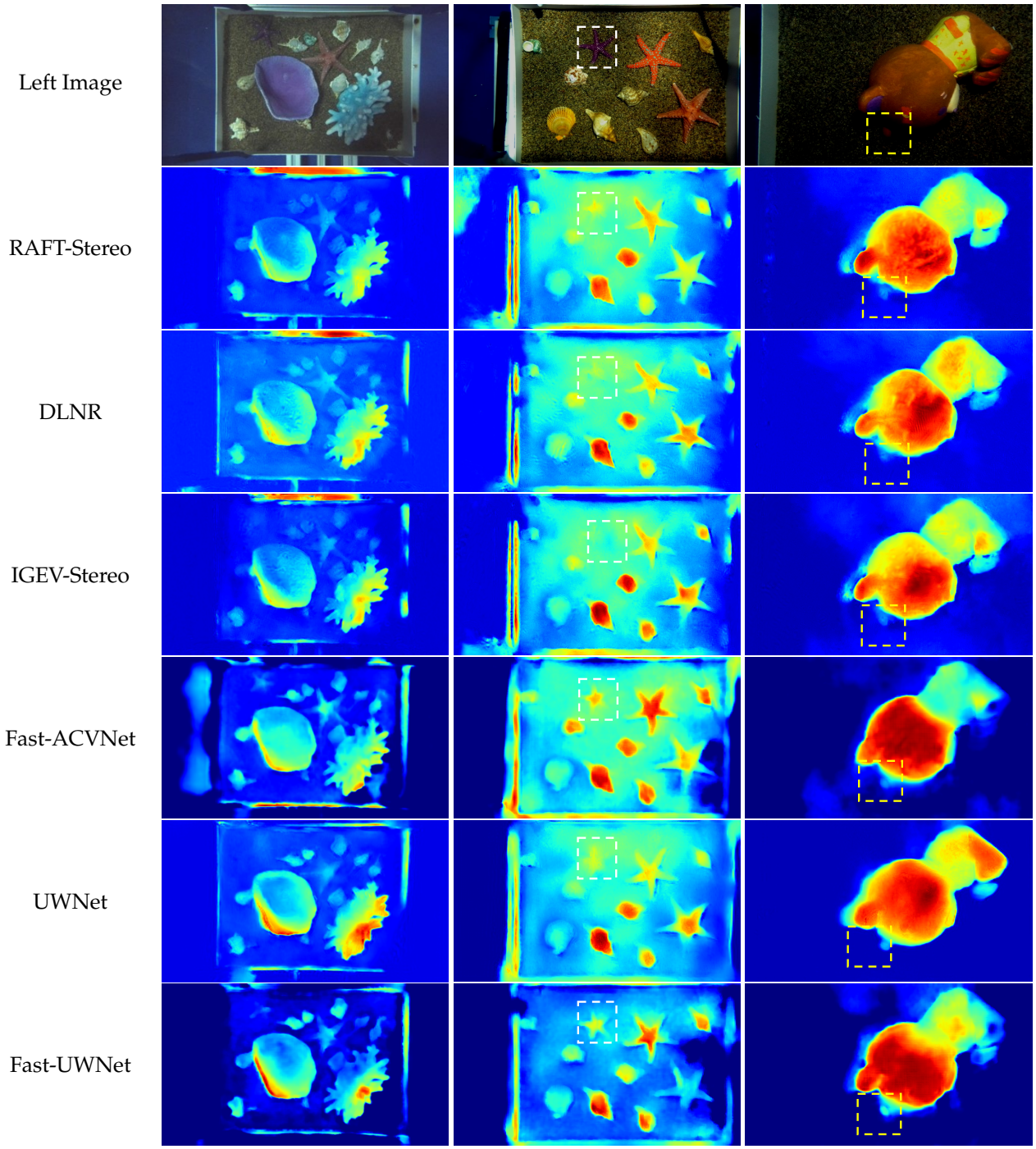


Figure 8. Visualization on the underwater dataset.

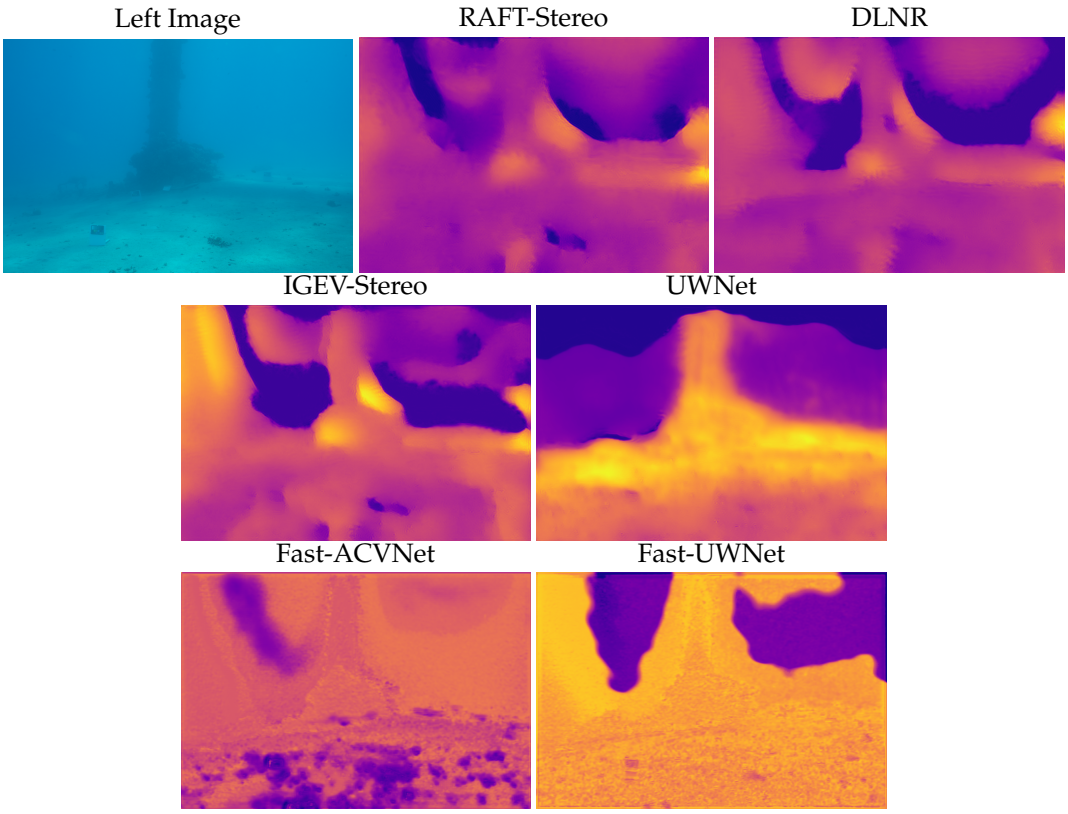


Figure 9. Visualization on the SQUID dataset.

On the Scene Flow dataset, The changes in the training process indicators are shown in Figure 10. As detailed in Table 7, UWNet achieved the best performance across all relevant metrics, particularly excelling in edge definition and clarity within the disparity maps, as visualized in Figure 11. Our method produced sharper, more defined edges, enhancing the accuracy of disparity estimation, especially in complex scenes with intricate structural elements. Compared to the lightweight model Fast-ACVNet [33], Fast-UWNet achieved better performance on both EPE and D1 metrics, demonstrating its effectiveness even within a constrained model structure.

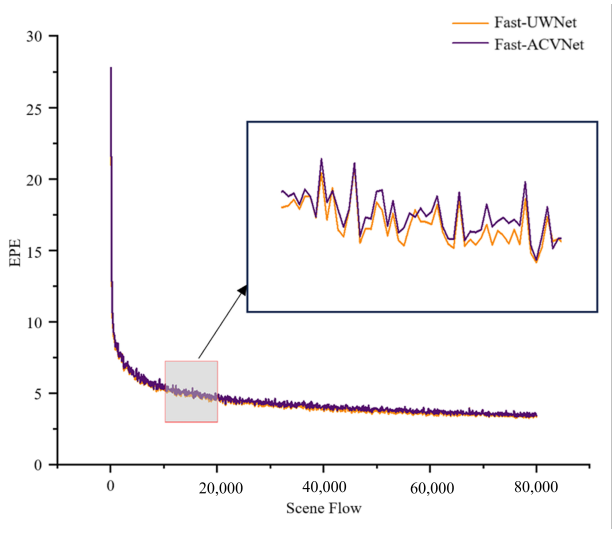


Figure 10. Performance comparison between Fast-ACVNet and Fast-UWNet during training shows a clear trend: our method consistently outperforms the other in terms of EPE.

Table 7. Quantitative evaluation on Scene Flow test set.

Method	RAFT [10]	DLNR [12]	IGEV [11]	Fast-ACV [33]	UWNet	Fast-UWNet
EPE (px)	2.43	2.39	2.63	2.45	**2.36**	**2.40**
>1px (%)	36.12	35.98	37.86	31.32	**35.71**	**31.02**
Time (ms)	132	297	180	49	270	**22**
Params (M)	11.23	57.37	12.62	5.22	12.05	3.30

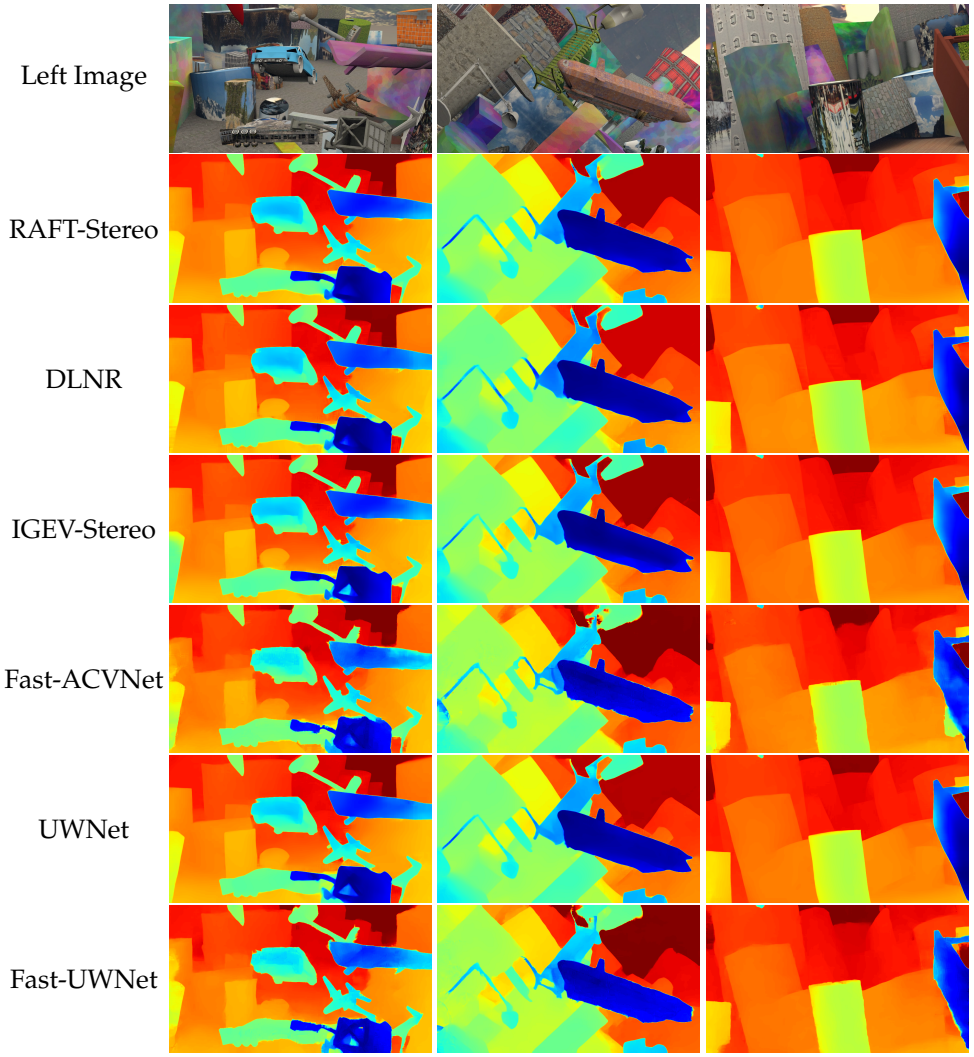


Figure 11. Visualization on the Scene Flow dataset.

4. Conclusions

In this work, we developed two stereo matching networks tailored for underwater environments: UWNet and its lightweight version, Fast-UWNet. UWNet addresses underwater imaging challenges, including noise, blurriness, and distortion, through self- and cross-attention mechanisms and an adaptive 1D-2D cross-search to improve cost volume representation. Additionally, a cascaded update module refines disparity estimation for more accurate depth mapping.

Fast-UWNet builds on this foundation with a streamlined design focused on efficiency, achieving fast inference at 0.02 s per image. By generating low-resolution disparity candidates directly and using confidence-based interpolation, it minimizes iterative updates, enabling precise disparity estimation with reduced computational complexity. Tests across diverse underwater settings demonstrate both networks' effectiveness in achiev-

ing reliable depth perception, marking a significant step forward in underwater stereo matching technology.

Author Contributions: Conceptualization, L.Z. and Y.G.; Methodology, L.Z.; Validation, J.Z.; Data curation, Y.L.; Writing—original draft preparation, L.Z.; Writing—review and editing, Y.G.; Visualization, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China, grant numbers 62401310 and 42406216; Natural Science Foundation of Shandong Province under Projects ZR2024QD031 and ZR2021QF028; and the opening project of Shandong Province Engineering Research Centre (Qingdao University of Science and Technology), grant numbers KF2024SD003 and KF2024SD009.

Data Availability Statement: The original data presented in the study are openly available in the following repositories: the Underwater Dataset at https://pan.baidu.com/s/12DTwuX3Ww0uXmHus\X6fN8A(pwd=4heh), accessed on 1 October 2023, the SQUID Dataset at https://csms.haifa.ac.il/profiles/tTreibitz/datasets/\ambient_forwardlooking/index.html, accessed on 1 September 2024 and the Scene Flow Dataset at https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html, accessed on 1 September 2023.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Rodionov, A.Y.; Dubrovin, F.; Unru, P.; Kulik, S.Y. Experimental research of distance estimation accuracy using underwater acoustic modems to provide navigation of underwater objects. In Proceedings of the 2017 24th Saint Petersburg International Conference on Integrated Navigation Systems (ICINS), Saint Petersburg, Russia, 29–31 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
2. Tan, C.S.; Mohd-Mokhtar, R.; Arshad, M.R. Fast fourier transform overlap approach for underwater acoustic positioning system. In Proceedings of the 2018 IEEE 8th International Conference on Underwater System Technology: Theory and Applications (USYS), Wuhan, China, 1–3 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–5.
3. Negre, P.L.; Bonin-Font, F.; Oliver, G. Cluster-based loop closing detection for underwater slam in feature-poor regions. In Proceedings of the 2016 IEEE international conference on robotics and automation (ICRA), Stockholm, Sweden, 16–21 May 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2589–2595.
4. Digumarti, S.T.; Chaurasia, G.; Taneja, A.; Siegwart, R.; Thomas, A.; Beardsley, P. Underwater 3d capture using a low-cost commercial depth camera. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–9.
5. Akkaynak, D.; Treibitz, T.; Shlesinger, T.; Loya, Y.; Tamir, R.; Iluz, D. What is the space of attenuation coefficients in underwater computer vision? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4931–4940.
6. Xu, Y.; Yu, D.; Ma, Y.; Li, Q.; Zhou, Y. Underwater stereo-matching algorithm based on belief propagation. *Signal, Image Video Process.* **2023**, *17*, 891–897. [CrossRef]
7. Ye, X.; Zhang, J.; Yuan, Y.; Xu, R.; Wang, Z.; Li, H. Underwater depth estimation via stereo adaptation networks. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5089–5101. [CrossRef]
8. Xu, B.; Xu, Y.; Yang, X.; Jia, W.; Guo, Y. Bilateral grid learning for stereo matching networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12497–12506.
9. Ye, X.; Li, Z.; Sun, B.; Wang, Z.; Xu, R.; Li, H.; Fan, X. Deep joint depth estimation and color correction from monocular underwater images based on unsupervised adaptation networks. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 3995–4008. [CrossRef]
10. Lipson, L.; Teed, Z.; Deng, J. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In Proceedings of the 2021 International Conference on 3D Vision (3DV), London, UK, 1–3 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 218–227.
11. Xu, G.; Wang, X.; Ding, X.; Yang, X. Iterative geometry encoding volume for stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 21919–21928.
12. Zhao, H.; Zhou, H.; Zhang, Y.; Chen, J.; Yang, Y.; Zhao, Y. High-frequency stereo matching network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1327–1336.
13. Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; Liu, S. Practical stereo matching via cascaded recurrent network with adaptive correlation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16263–16272.

14. Chen, Z.; Long, W.; Yao, H.; Zhang, Y.; Wang, B.; Qin, Y.; Wu, J. Mocha-stereo: Motif channel attention network for stereo matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 27768–27777.
15. Shi, Y. Rethinking iterative stereo matching from diffusion bridge model perspective. *arXiv* **2024**, arXiv:2404.09051.
16. Xiao, W.; Zhao, W. Rectified iterative disparity for stereo matching. *arXiv* **2024**, arXiv:2406.10943.
17. Ubina, N.A.; Cheng, S.-C.; Chang, C.-C.; Cai, S.-Y.; Lan, H.-Y.; Lu, H.-Y. Intelligent underwater stereo camera design for fish metric estimation using reliable object matching. *IEEE Access* **2022**, *10*, 74605–74619. [CrossRef]
18. Berman, D.; Levy, D.; Avidan, S.; Treibitz, T. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2822–2837. [CrossRef] [PubMed]
19. Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4040–4048.
20. Wang, H.; Sun, S.; Bai, X.; Wang, J.; Ren, P. A reinforcement learning paradigm of configuring visual enhancement for object detection in underwater scenes. *IEEE J. Ocean. Eng.* **2023**, *48*, 443–461. [CrossRef]
21. Wang, H.; Zhang, W.; Bai, L.; Ren, P. Metalantis: A comprehensive underwater image enhancement framework. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5618319. [CrossRef]
22. Wang, H.; Sun, S.; Chang, L.; Li, H.; Zhang, W.; Frery, A.C.; Ren, P. Inspiration: A reinforcement learning-based human visual perception-driven image enhancement paradigm for underwater scenes. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108411. [CrossRef]
23. Wang, H.; Zhang, W.; Ren, P. Self-organized underwater image enhancement. *ISPRS J. Photogramm. Remote Sens.* **2024**, *215*, 1–14. [CrossRef]
24. Huang, X.; Liu, M.-Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 172–189.
25. Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; Yang, M.-H. Diverse image-to-image translation via disentangled representations. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 35–51.
26. Xu, Y.; Hou, J.; Zhu, X.; Wang, C.; Shi, H.; Wang, J.; Li, Y.; Ren, P. Hyperspectral image super-resolution with convlstm skip-connections. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5519016. [CrossRef]
27. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
28. Liu, M.-Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; Kautz, J. Few-shot unsupervised image-to-image translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10551–10560.
29. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6924–6932.
30. Simonyan, K. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
31. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
32. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. Loftr: Detector-free local feature matching with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8922–8931.
33. Xu, G.; Wang, Y.; Cheng, J.; Tang, J.; Yang, X. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 2461–2474. [CrossRef] [PubMed]
34. Liu, B.; Yu, H.; Long, Y. Local similarity pattern and cost self-reassembling for deep stereo matching networks. In Proceedings of the 36th AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 1647–1655.
35. Gan, Y.; Xu, X.; Sun, W.; Lin, L. Monocular depth estimation with affinity, vertical pooling, and label enhancement. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 224–239.

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.