Journal of *Marine Science and Engineering*

MDPI

Article

A Semantic Segmentation Method Based on Image Entropy Weighted Spatio-Temporal Fusion for Blade Attachment Recognition of Marine Current Turbines

Fei Qi and Tianzhen Wang *

Logistics Engineering College, Shanghai Maritime University, Shanghai 201306, China; 202030210045@stu.shmtu.edu.cn

* Correspondence: tzwang@shmtu.edu.cn

Abstract: Marine current turbines (MCTs) may exhibit reduced energy production and structural instability due to attachments, such as biofouling and plankton. Semantic segmentation (SS) is utilized to recognize these attachments, enabling on-demand maintenance towards optimizing power generation efficiency and minimizing maintenance costs. However, the degree of motion blur might vary according to the MCT rotational speed. The SS methods are not robust against such variations, and the recognition accuracy could be significantly reduced. In order to alleviate this problem, the SS method is proposed based on image entropy weighted spatio-temporal fusion (IEWSTF). The method has two features: (1) A spatio-temporal fusion (STF) mechanism is proposed to learn spatio-temporal (ST) features in adjacent frames while conducting feature fusion, thus reducing the impact of motion blur on feature extraction. (2) An image entropy weighting (IEW) mechanism is proposed to adjust the fusion weights adaptively for better fusion effects. The experimental results demonstrate that the proposed method achieves superior recognition performance with MCT datasets with various rotational speeds and is more robust to rotational speed variations than other methods.

Keywords: marine current turbine; vision transformer; semantic segmentation; blade attachment; image entropy

Citation: Qi, F.; Wang, T. A Semantic Segmentation Method Based on Image Entropy Weighted Spatio-Temporal Fusion for Blade Attachment Recognition of Marine Current Turbines. *J. Mar. Sci. Eng.* **2023**, *11*, 691. https://doi.org/10.3390/jmse11040691

Academic Editors: Mohamed Benbouzid, Anne Blavette and Nicolas Guillou

Received: 24 February 2023
Revised: 16 March 2023
Accepted: 22 March 2023
Published: 24 March 2023

1. Introduction

As climate change becomes more intense, the development of clean energy sources is gradually accelerating [1,2]. Marine current energy is recognized as a promising source of clean energy due to its predictability, high energy density and limited environmental impact [3,4]. The marine current turbine (MCT) can convert marine current energy into electrical energy. However, MCTs are deployed underwater, and attachments can accumulate over time on the submerged surfaces of the MCT, such as plankton or biofouling [5]. The work of Farkas et al. [6] indicates that more than 80% of the areas on the blades of a normally operating MCT have the potential to be attached. Attachments can cause increased fatigue load and unbalanced torque, which can threaten the structural stability of the MCT. Most importantly, there is a reduction in the energy conversion efficiency of MCTs since the roughness of the blade surface is changed by attachment [6–8]. Therefore, it is necessary to clean and maintain the MCT to ensure its stable and efficient operation. However, the marine environment poses problems in accessibility, which means that the maintenance of the MCT will be time-consuming and costly [9]. Consequently, it is important to recognize the attachment of MCT blades accurately and quickly, which facilitates the optimization of maintenance schedules and even on-demand maintenance, thus reducing maintenance and monitoring costs and improving power generation efficiency.

Recent research has focused on developing methods for recognizing attachments on MCT blades, which can be classified into two categories, as summarized by Xie et al. [5]. The first category involves extracting features from the electrical signals obtained from the

built-in sensors of the MCT to recognize attachment, which has been demonstrated to be fast and effective [10–12]. However, these methods are likely to respond only when there is unbalanced torque or significant speed variations caused by the attachment. Additionally, the extracted features may be submerged due to strong turbulence interference, which may lead to recognition failure. In contrast, image signals are less affected by turbulence since they contain rich spatial information. Thus, the second category involves utilizing image signals from optical or image sensors to recognize attachments. Convolutional neural networks (CNNs) have been widely used to extract image features for recognizing attachments on MCT blades. Typically, a CNN encoder is used for feature extraction, which is followed by a corresponding classifier or decoder to achieve recognition. Zheng et al. [13] proposed a method that utilizes a sparse auto-encoder and softmax regression for the feature extraction and classification of various attachment cases through MCT images. However, this method has a complex training process and can only roughly classify a few pre-defined cases. In comparison, Xin et al. [14] utilized depthwise separable convolutions [15] instead of sparse auto-encoders to classify more complex attachment cases. However, the above methods may fail to recognize the attachment cases outside of the pre-defined standards, and it is challenging to obtain the location or coverage of the attachment exactly.

In contrast to the previous methods, semantic segmentation (SS) is capable of classifying each pixel of an image, which can eliminate the drawbacks associated with the above methods. Peng et al. [16] first introduced SS to the attachment recognition of MCT blades, combining fine features into the encoder to optimize the segmentation contours and improve accuracy. The pixel-level recognition results can be used as output and visualized, thus giving detailed information on the location and coverage of the attachment and therefore facilitating the provision of real-time and detailed information on the status of the blades to the monitoring personnel. However, the MCT operation is affected by various factors, such as turbulence, and the rotational speed is not constant. These factors result in different degrees of motion blur in the acquired MCT images, which causes the degradation of feature extraction, leading to changes in the edge and texture features and potential recognition failures [17]. This will lead to a significant reduction in the recognition accuracy of SS when the rotational speed of the MCT changes. The above problem shows the poor robustness of the SS method against the variation of the MCT rotational speed, which leads to the inability to achieve the ideal recognition performance when encountering an unknown rotational speed situation. In order to address this issue, a SS method is proposed for the attachment recognition of MCT blades in this paper, which is based on image entropy weighted spatio-temporal fusion (IEWSTF). The core idea of this method is to perform feature extraction on multiple adjacent frames simultaneously and construct a feature mapping from other frames to the key frame through feature learning, which reflects the spatio-temporal (ST) features between adjacent frames and alleviates the degradation of feature extraction due to motion blur. A spatio-temporal fusion (STF) mechanism is designed to learn the ST features. In addition, an image entropy weighting (IEW) mechanism is proposed to adaptively adjust the fusion weights of adjacent frames to prevent inaccurate feature learning due to the excessive differences between frames. The experimental results demonstrate the superior performance of the proposed method compared to other methods, with highly accurate attachment recognition for MCT blades at high rotation speeds and robustness against rotational speed variations.

The main work of this paper is summarized as follows:

1. An IEWSTF-based semantic segmentation method is proposed in this paper for the attachment recognition of MCT blades, which aims to improve robustness against the variations in rotational speed of the MCT;
2. The STF mechanism is proposed for learning the ST features between the adjacent frames to alleviate the degradation of feature extraction due to motion blur;
3. An IEW mechanism is proposed to obtain the optimal fusion features by adaptively adjusting the fusion weights by measuring the degree of difference between the adjacent frames.

The remainder of this paper is arranged as follows. Section 2 provides a concise overview of prior research related to the present study, outlining their respective limitations in the context of attachment recognition on MCT blades. Section 3 introduces the MCT image dataset employed in the experiments in this paper and then presents a detailed exposition of the IEWSTF-based SS method proposed for robust attachment recognition. Section 4 presents the experimental results and discussion. Finally, Section 5 summarizes the main findings in this paper and provides some directions for future research.

2. Related Works

2.1. Semantic Segmentation

The SS techniques aim to predict the categorical assignment for each pixel in an image, thereby facilitating the depiction of their spatial interrelationships. The advent of the fully convolutional network (FCN) [18] revolutionized the application of deep learning models in SS by eliminating fully connected layers and introducing skip connections, enabling pixel-level prediction. However, the limited receptive field of the FCN hinders its ability to produce high-quality segmentation. To address this challenge, Chen et al. [19] introduced the concept of dilated convolution. In its subsequent iterations [20–22], a pyramid structure is established through the incorporation of dilated convolution with varying dilation rates, thereby expanding the receptive field and facilitating the modeling of contextual information. However, these advancements have failed to fundamentally address the limitation of the fixed receptive field size on CNNs, thereby impeding further progress in the field of SS. Furthermore, too many empirical modules have been introduced to improve the segmentation performance [23], which leads to computational complexity and makes it unsatisfactory for the processing speed when used for the attachment recognition of MCTs.

In recent times, transformer and self-attention models have garnered significant attention in computer vision, with experiments demonstrating the superiority of vision transformers (ViTs) with their variable receptive fields over CNNs [24,25]. There have been extensive works developing ViT-based SS methods that still follow the encoder-decoder structure that is generally adopted. Earlier efforts include SETR [26] and TransUNet [27], both of which attempt to replace the CNN in the encoder with the ViT for early-stage feature extraction. It has been observed that the ViT has the ability to capture global contextual relationships from the outset, thereby overcoming the drawback of the limited receptive field of CNNs. Yet, the difference is that SETR employs the ViT exclusively in its encoder, while TransUNet utilizes a hybrid structure comprising a variant CNN and the ViT. However, the above methods fail to entirely alleviate the limitations of the CNN, as their decoders are still constructed by convolutional layers. In response to this problem, convolutional layers are replaced with linear layers in the decoder by SegFormer [23], thus constructing the SS framework completely without convolution. The framework realizes high efficiency through the complete utilization of linear layers. Meanwhile, excellent segmentation performance is maintained by the great advantage of ViTs in extracting global contextual information.

To the best of our knowledge, limited research has been conducted to evaluate the performance of ViT-based models in recognition of MCT blade attachments. In this paper, we will evaluate the performance of several state-of-the-art ViT-based SS methods for MCT attachment recognition and propose a new SS method to enhance the robustness against rotational speed variations.

Meanwhile, as motion blur remains a challenge in the SS field, many efforts have attempted to overcome this problem. Li et al. [28] improved the accuracy of SS for motion-blurred images by fusing the features of ViT and convolutional layers. Some works [29–32] optimize the semantic segmentation performance with the help of relevant information from history frames to overcome motion blur. However, it is difficult to achieve an optimal estimation of the exact position of motion-blurred objects with the information of history frames only. Wang et al. [33] used the information in the next adjacent frame to overcome the effect of motion blur. However, this method requires the help of optical flow estimation

to model the motion trend of the object, which will create additional computational costs and also require additional a priori knowledge for support. The goal of this paper is to utilize the spatio-temporal information contained in both the historical and future frames, thus filling the gap in the above research. Moreover, such information is fused into the feature maps to better estimate the position of objects with motion blur in the current frame. In addition, by aggregating multi-scale features, it is able to optimize the segmentation edges and achieve optimal attachment recognition.

2.2. Image Entropy

In the field of thermodynamics, entropy is utilized to characterize the disorder or randomness of molecular motion. Shannon introduced the concept of entropy to quantify the uncertainty of an information source in communication systems. The information entropy can be interpreted as the amount of information present in a source, with higher entropy indicating a greater amount of information. As such, it is frequently utilized as an objective function in optimization problems.

In the field of image processing, various entropies are utilized for the determination of threshold values for image segmentation [34–36], where the optimal threshold is determined by maximizing the entropy value. However, in this paper, entropy is employed to measure the degree of difference between adjacent frames, which serves as a reference for weight assignment and thus helps to mitigate the distortion of feature fusion caused by excessive differences between adjacent frames. A detailed analysis is presented in Section 3.2.2.

3. Data and Methods

3.1. Image Dataset for MCT

For the purpose of model training, the MCT data is sourced from the marine environment simulation and experimental platform of the marine current generator set at Shanghai Maritime University, as depicted in [16]. In order to simulate the operation of the MCT on the seafloor, an MCT prototype is placed in a water tank, which is used to replicate the flow of water. The pump is used to propel the water flow, thereby inducing the rotation of the MCT. An underwater camera is placed in the water tank to record a frontal video of the MCT prototype at a resolution of 640 $\times$ 576. The original images are acquired through video frame capturing and resized to 480 pixels in width and height.

The blade attachment occurs through the growth of biofilms on the submerged surface of the MCT and the accumulation of plankton by adsorption [6]. However, it is not practical to cultivate biofilms on the prototype or simulate plankton accumulation due to the time-consuming nature of the process, the uncontrollability of the results, and the potential for irreversible damage to the prototype. Therefore, ropes are wrapped around the blades to simulate the textural features of the attachment in this paper. Considering that the blade is less likely to be attached to the locations that are closer to the tip [6], the rope is wrapped from the root of the blade while leaving the tip of the blade empty. In order to prevent uniform textural features, variations are introduced, including the use of ropes of different colors and forms of wrapping, etc. In summary, five forms of attachment are simulated. Figure 1 shows examples of the five simulations and their corresponding hand-labeled ground truth (GT). The GT utilizes gray to represent the healthy part of the blade; red indicates the attachment and light blue indicates the shaft that carries the blade.

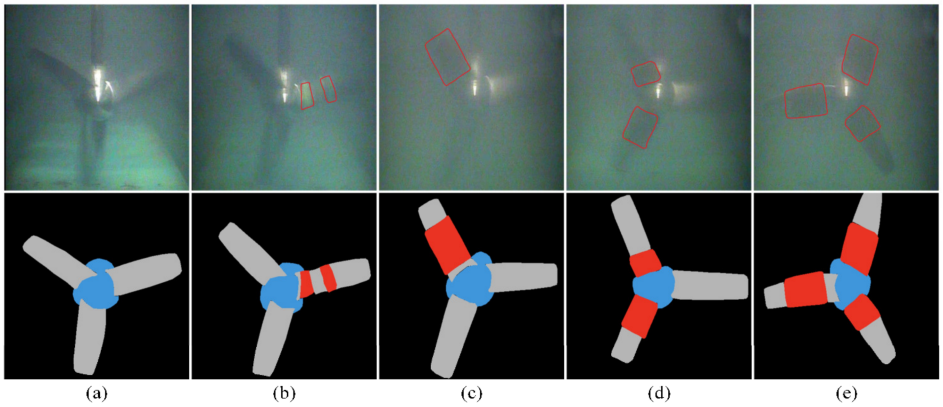

Figure 1. The images of MCT blades and the corresponding GT under different conditions of attachment simulation. The red box in the images indicates the area covered by the attachment. The forms of attachment are: (**a**) healthy; (**b**) single blade sparsely attached; (**c**) single blade densely attached; (**d**) double blades densely attached; (**e**) triple blades densely attached.

In order to evaluate the effect of the rotational speed on SS accuracy, it is necessary to acquire MCT image data at different rotational speeds. In fact, the degree of motion blur determines the effect of the rotational speed on SS accuracy. Therefore, it is necessary to find the relationship between the rotational speed of the MCT and the degree of motion blur so as to determine the rotational speed reasonably during the image acquisition. There are two factors that determine the degree of motion blur: one is the exposure time of the image acquisition device, and the other is the displacement of the object in the frame that occurs during the exposure time. With a certain exposure time, motion blur is mainly related to the latter factor, i.e., the rotational speed of the MCT. Since the MCT prototype lacks a speed sensing module, the rotational speed is measured through the stator current frequency output from the prototype's permanent magnet synchronous generator (PMSG). The reason is that, in the PMSG, the shaft rotation frequency has the following relationship with the stator current frequency [10]:

$$f_m = \frac{f_e}{p} \tag{1}$$

where f_e denotes the frequency of the stator current, the p is the pole pair of the PMSG, and f_m represents the rotational frequency of the shaft.

Therefore, f_e and p can be used to determine the rotational speed of the MCT blades. Further, it is necessary to derive the relationship between the rotational speed of the MCT and motion blur. The MCT turns through an angle $\Delta\theta$ during the exposure time t_{exp} of the image acquisition device, resulting in motion blur. By combining with Equation (1), it is obtained that

$$\Delta\theta = \frac{f_e \cdot t_{exp}}{p} \times 360^\circ \tag{2}$$

where t_{exp} denotes the exposure time of the image acquisition device. In the experimental platform introduced in the previous section, $t_{exp} = 1/90$ s and $p = 8$.

The parameter $\Delta\theta$ is appropriate to reflect the degree of motion blur of the MCT. As a result, this paper divides five levels of rotational speed according to the $\Delta\theta$ of the MCT: level 1 for $\Delta\theta \leq 6^\circ$, level 2 for $6^\circ < \Delta\theta \leq 7^\circ$, level 3 for $7^\circ < \Delta\theta \leq 8^\circ$, level 4 for $8^\circ < \Delta\theta \leq 9^\circ$, and level 5 for $\Delta\theta > 9^\circ$. Thus, the determined ranges of the stator current frequencies are $f_e \leq 12$ Hz, 12 Hz $< f_e \leq 14$ Hz, 14 Hz $< f_e \leq 16$ Hz, 16 Hz $< f_e \leq 18$ Hz, and $f_e > 18$ Hz. The MCT images will be acquired in each of these five current frequency ranges. Similarly, images are acquired for five attachment types at each speed level, as shown in Figure 1.

Data augmentation is applied to enrich the dataset further to enhance the generalization and robustness of the model. Two approaches are used in this paper: random rotation and photometric distortion. Random rotation is used to rotate the original image and GT by the same random angles. Since the MCT is rotating, this approach is able to simulate various spatial positions of the MCT blades while reducing the effort of processing raw data. Photometric distortion is used to randomize the hue, contrast and saturation of the image during the training phase, thus avoiding over-fitting due to the model's over-reliance on color and background features. There are 48 images for each attachment form at each level of rotational speed, totaling 1200 images in the dataset.

3.2. The Image Entropy Weighted Spatio-Temporal Fusion-Based SS Method

Figure 2 illustrates the IEWSTF-based SS method proposed in this paper. The network structure of the proposed method still follows the basic structure of the encoder-decoder. The ViT is employed in the encoder part of the proposed method to ensure a sufficiently large perceptual field for the model and to enable a fast SS. The encoder contains four stages, each of which will reduce the resolution of the feature maps by half so that the multi-scale features are extracted. In the encoder, the STF mechanism is designed to learn the ST features contained between the adjacent frames. Different features are learned and extracted from the three adjacent frames $\boldsymbol{F}(t)$, $\boldsymbol{F}(t-1)$ and $\boldsymbol{F}(t+1)$, respectively. These feature maps will be used for fusion. More details will be described in Section 3.2.1. In the above three frames, $\boldsymbol{F}(t)$ is used as the key frame, and GT is labeled according to $\boldsymbol{F}(t)$. The other two are used as reference frames. The network constructs a mapping from these reference frames to the key frame features, which reflects the changes that occur in the reference frames relative to the key frame, i.e., the ST features.

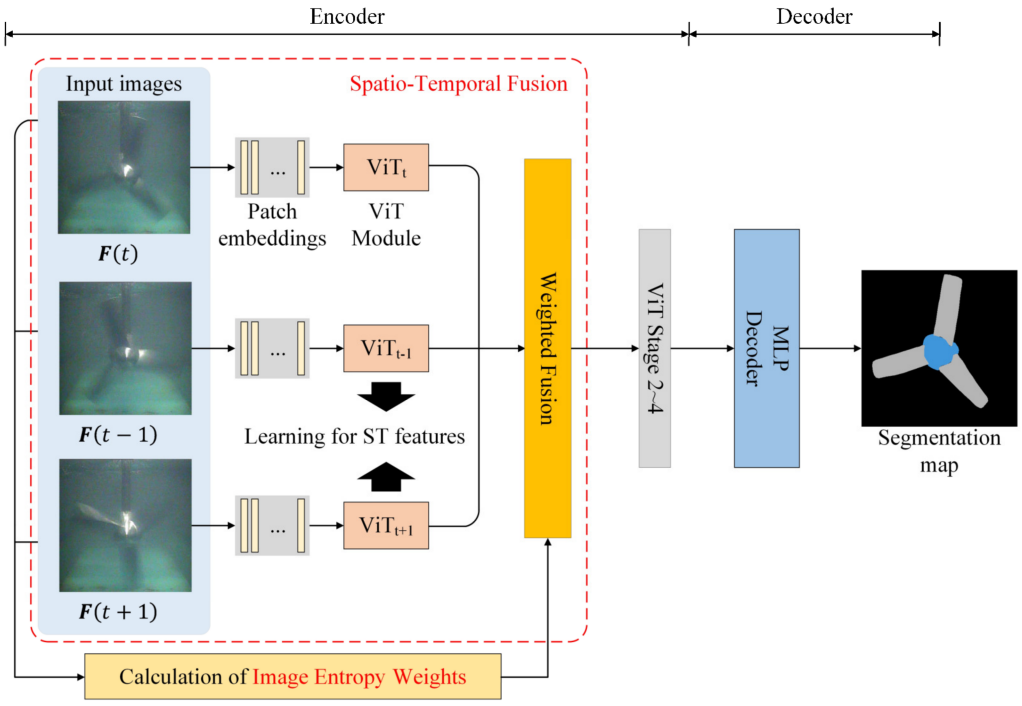

Figure 2. The structure diagram of the proposed IEWSTF-based SS method.

In the fusion process, an IEW mechanism is proposed in order to avoid distortion of the fused features caused by the excessive differences between the adjacent frames. Smaller weights will be assigned to the features that differ excessively, thus weakening the distortion of the fused features. A description of the proposed IEW mechanism will be given in Section 3.2.2.

In the decoder of the method, a lightweight structure is constructed based on the multilayer perceptron (MLP) that will accept multilevel features from the encoder as input [23]. Through the fusion of multilevel features, both global and local features are preserved, thus allowing for high-quality segmentation results with fewer computations.

In order to further reduce the model computations and thus increase the speed of attachment recognition, the segmentation map size output by the model is set to be smaller than the input image. When the input size is $(h_i,\ w_i)$, the size of the segmentation map output by the model is $(h_i/4,\ w_i/4)$. The bilinear interpolation upsampling is used to scale up the segmentation map to (h_i, w_i) in order to compare the segmentation results with the GT of size $(h_i,\ w_i)$ for evaluating the performance.

3.2.1. Spatio-Temporal Fusion

In order to learn the ST features contained in the adjacent frames for the network, the first step is to calculate the embeddings of the adjacent frames $\boldsymbol{F}(t)$, $\boldsymbol{F}(t-1)$ and $\boldsymbol{F}(t+1)$ for feeding them into ViTs for feature learning. For effective feature learning, on the one hand, there need to be some differences between the features to be fused. Excessively small differences can cause the weights that are learned by the three ViTs to converge to the same, and therefore, no ST features will be learned. On the other hand, the differences between the adjacent frames are mainly reflected in the variations in the spatial positions of the objects in the images. For the MCT images, the deep features in the network tend to reflect the abstract and global features with little difference. In contrast, the shallow features will retain more contour and texture features and are more likely to highlight the differences between the adjacent frames. For the above reasons, the learning and fusion of the ST features are done in the shallow layers of the encoder. Specifically, in Stage 1 of the encoder, three ViT modules are used to learn the features in $\boldsymbol{F}(t)$, $\boldsymbol{F}(t-1)$ and $\boldsymbol{F}(t+1)$, respectively. The three sets of embeddings will be fed into the corresponding ViT module for feature learning, respectively, which has been computed in the first step. As depicted in Figure 2, $\mathrm{ViT_t}$ will learn the features of the key frame $\boldsymbol{F}(t)$ by iteratively updating the network parameters. The feature maps will be used as the foundation for SS. Furthermore, since the ground truth for segmentation is annotated based on the key frame $\boldsymbol{F}(t)$, the feature maps extracted by $\mathrm{ViT_{t-1}}$ and $\mathrm{ViT_{t+1}}$ in the adjacent frames $\boldsymbol{F}(t-1)$ and $\boldsymbol{F}(t+1)$ will also gradually approach the feature maps corresponding to the key frame $\boldsymbol{F}(t)$ during the iterative update of the network parameters. In fact, this will enable $\mathrm{ViT_{t-1}}$ and $\mathrm{ViT_{t+1}}$ to construct mappings from the adjacent frames to the key frame features, which can be considered as the network capturing the ST features of the adjacent frames [29,37]. It is worth noting that the above ViT modules use a similar structure to the Transformer Block in SegFormer. However, this paper uses three ViT modules to process three frames, respectively. The additional ViT modules are used to learn the spatio-temporal features in the adjacent frames and fuse them with the features of the key frames. The process of STF can be formulated as Equation (3):

$$\boldsymbol{G}[\boldsymbol{F}(t), \boldsymbol{F}(t-1), \boldsymbol{F}(t+1)] = \boldsymbol{W}_t \boldsymbol{F}(t) + w_{\mathrm{IE}}[\boldsymbol{W}_{t-1}\boldsymbol{F}(t-1) + \boldsymbol{W}_{t+1}\boldsymbol{F}(t+1)] \tag{3}$$

where $\boldsymbol{G}[\boldsymbol{F}(t), \boldsymbol{F}(t-1), \boldsymbol{F}(t+1)]$ are the fused features. The w_{IE} is the image entropy weights, which will be described in Section 3.2.2. $\boldsymbol{W}_t$, $\boldsymbol{W}_{t-1}$ and $\boldsymbol{W}_{t+1}$ denote the weight matrix of $\mathrm{ViT_t}$, $\mathrm{ViT_{t-1}}$ and $\mathrm{ViT_{t+1}}$, respectively.

3.2.2. Image Entropy Weighting

It is necessary to fuse the feature maps of the adjacent frames described above to improve the robustness of the SS network against the variations in rotational speed. To avoid feature fusion distortion due to the excessive differences between adjacent frames, an IEW mechanism is proposed to adjust the values of the fusion weights adaptively. Lower fusion weights will be assigned to the features of the adjacent frames with larger differences. Specifically, the difference matrix is calculated between adjacent frames. Then, the image entropy is obtained from the difference matrix, and then the fusion weights are calculated.

For frame $\boldsymbol{F}(t)$, it is subtracted from the neighboring frames $\boldsymbol{F}(t-1)$ and $\boldsymbol{F}(t+1)$, respectively, to obtain the corresponding difference matrix $\boldsymbol{D}_{t-1}$ and $\boldsymbol{D}_{t+1}$:

$$\begin{cases} \boldsymbol{D}_{t-1} = \text{Squeeze}(\text{ReLU}(\boldsymbol{F}(t) - \boldsymbol{F}(t-1)))/(255 \times 3) \\ \boldsymbol{D}_{t+1} = \text{Squeeze}(\text{ReLU}(\boldsymbol{F}(t) - \boldsymbol{F}(t+1)))/(255 \times 3) \end{cases} \tag{4}$$

where ReLU denotes the rectified linear unit (ReLU). The ReLU filters out values less than zero so as to ensure that the information retained in the difference matrix is based on the $F(t)$. The matrix continues three channels after the ReLU, which is the same as an image. The Squeeze represents the squeezing of three channels into one, i.e., summing up the values of each pixel over the three channels. Since the image pixels have a maximum value of 255 on each channel, the Squeeze sums the three channels so that the maximum value becomes 255×3. For normalizing all the values, the matrix is divided by 255×3.

However, there is considerable noise in the difference matrix calculated directly from the original image, which is presented as discrete noise in the background region. In contrast, it is continuous in the region belonging to the MCT blade in the difference matrix, which is fundamentally different from the background. Utilizing this characteristic, convolve filtering is used to perform background denoising on the difference matrix. It is an effective algorithm to reduce the pixel value of the noise by filtering the local area with its local average. The kernel size of the convolving filter is set to $(7,7)$. After convolve filtering, the noise removal is then achieved by setting the values to zero for those less than the threshold τ in the difference matrix. For $\boldsymbol{D}_{t+1}$, the above process can be described as Equation (5):

$$\boldsymbol{D'}_{t+1} = \text{ReLU}(\text{convolve}_{7\times7}(\boldsymbol{D}_{t+1}) - \tau). \tag{5}$$

where $\text{convolve}_{7\times7}$ denotes the convolve filtering with the kernel size of $(7,7)$, and $\boldsymbol{D'}_{t+1}$ is the difference matrix after denoising.

Equation (5) will be repeated n times to achieve the optimal denoising effect. The denoising of $\boldsymbol{D}_{t-1}$ is done in the same way. For the image data of the MCT in this paper, it is set as $\tau = 0.55$ and $n = 4$.

Then, for the denoised difference matrices $\boldsymbol{D'}_{t-1}$ and $\boldsymbol{D'}_{t+1}$ above, their image entropies are calculated according to Equation (6) [38]:

$$h(f) = -\sum_{i=0}^{255} p_i \log_2 p_i \tag{6}$$

where p_i denotes the probability of the appearance of the pixels with gray value i in the image f.

To calculate the entropy of the difference matrix, all values in the difference matrix $\boldsymbol{D'}_{t-1}$ and $\boldsymbol{D'}_{t+1}$ are multiplied by 255 and rounded down; thus, the difference matrix can be considered a grayscale map to calculate the image entropy. Further, the normalized results are obtained after the calculation of the sigmoid function in order to be used in the weight calculations. The operation above is shown in Equation (7):

$$h_{\text{normal}} = \sigma\big(h\big(\text{floor}\big(\boldsymbol{D'} \times 255\big)\big)\big) \tag{7}$$

where σ denotes the sigmoid function, and the floor indicates rounding down.

At this point, the weight values are obtained via Equation (8), which will be assigned to the feature maps:

$$w_{\mathbf{IE}} = 1 - h_{\text{normal}} \tag{8}$$

where $w_{\mathbf{IE}}$ denotes the weight that will be assigned to the feature maps of adjacent frames.

When there is a larger gap between the adjacent frames, there will be a larger scatter of points in the difference matrix indicating the MCT blades, which means a higher entropy value, i.e., a larger value of h_{normal}. After the operation of Equation (8), the weights assigned to these features will be smaller. The above process is the IEW proposed in this paper. The

IEW adaptively adjusts the fusion weights according to the differences in adjacent frames, thus weakening the distortion of feature fusion by the excessive differences.

4. Results and Discussion

A comprehensive comparative experiment was conducted on the MCT dataset to evaluate the effectiveness of the proposed method. Section 4.1 outlines the experimental parameters of this paper, and Section 4.2 provides the evaluation metrics of the experiments. The evaluations are conducted in Sections 4.3 and 4.4, with different experimental condition settings, to assess the performance of the proposed method.

4.1. Configuration of the Training Process

In this paper, the experimental hardware platform is a single NVIDIA Quadro P5000 GPU with 15 GB of RAM. The software framework is built based on Pytorch. The parameters are set for the training process, as shown in Table 1. The initial learning rate is set to 0.001 using the AdamW optimizer, and the batch size is set to 4 due to the fixed input image size of 480 $\times$ 480 and the memory limitation of the hardware platform. All models are trained from scratch, and the weight parameters are randomly initialized using the truncated normal distribution.

Table 1. Experimental parameters.

Parameters	Value
Optimizer	AdamW
Batch size	4
Initial learning rate	0.001
Epochs	100
Shuffle	True
Photometric Distortion	True

During the training process, data augmentation is used to increase the richness of the dataset to ensure that the trained models have sufficient generalization performance, which has been introduced in Section 3.1. Further, all data are normalized to accelerate network convergence before being fed into the model.

4.2. Evaluation Metrics

The mean intersection-over-union (mIoU) is used to evaluate the SS accuracy of these approaches, which is the most commonly used evaluation metric for SS. In addition, pixel accuracy (PA) is also used in this paper to evaluate the recognition accuracy of the model in each semantic category in detail. The calculation of these two metrics has been described in [39].

In order to ensure a fair and reasonable comparison of the proposed method with other methods, this paper conducts two types of comparisons. The first comparison involves the use of all MCT image data, collected at five levels of rotational speed for training and testing, where 1200 images are randomly split into the training set and the test set with a 9:1 ratio. Section 4.3 presents the results and analysis obtained using this approach. The second comparison involves training the model using only a portion of the level 1 data. The 240 images from level 1 are randomly split into the training set and test set with a 9:1 ratio, and all 960 images from the other four speed levels (levels 2 to 5) are used for testing. Section 4.4 presents the comparison results and analysis for this approach.

4.3. Overall Accuracy Evaluation of SS on MCT Dataset

The proposed method is compared with the current mainstream ViT-based SS methods, including SETR and SegFormer, in which the performances are close to state-of-the-art. For SETR, the parameter setting follows the SETR-Naive-Base, which is the lightest configuration of the model. In terms of SegFormer, SegFormer-MiT-B0 is used, which is also the

smallest model configuration of SegFormer, for the consideration of the requirements for fast recognition.

Table 2 shows the comparison of the mIoU among these methods. It can be seen from Table 2 that a higher segmentation accuracy is achieved by the IEWSTF-based SS method proposed in this paper than the other two methods, with improvement exceeding 2% of the mIoU compared to the SETR model. The performance boost is even more than 4% compared to SegFormer. This is due to the fact that the proposed IEWSTF is able to perform efficient feature learning based on the adjacent frames, and the mappings from images to ST features can be constructed, thus overcoming the feature deviations caused by motion blur. Therefore, the proposed method achieves more accurate recognition of the MCT images at various speeds than other methods, resulting in a higher overall mIoU.

Table 2. Comparison of mIoU for all methods.

Methods	mIoU
SegFormer-MiT-B0	92.95
SETR-Naive-Base	94.26
IEWSTF (Ours)	**96.99**

To evaluate the recognition accuracy on each semantic category (blade, attachment, shaft) for all methods in more detail, the PA is calculated for each category, respectively. The comparison for PA is shown in Figure 3.

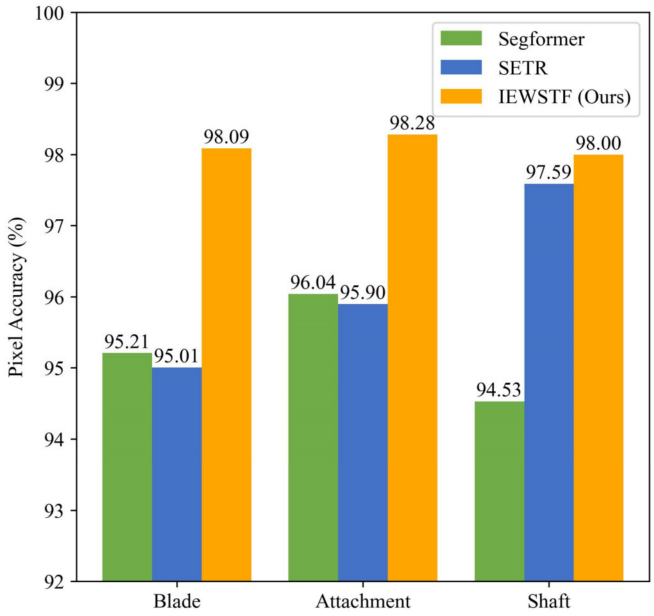

Figure 3. Comparison of PA for all methods in each semantic category.

As seen in Figure 3, the proposed method achieves 98% PA in all three categories, which is higher than the other two comparison methods. Furthermore, there is a significant improvement in PA in both the blade and attachment categories with respect to SETR and SegFormer, in the range of 2–3%, which shows the superiority of the proposed method in the attachment recognition of MCT blades. In fact, the motion blur is mainly reflected in the positions of the blades and attachments in the images, which indicates that the proposed IEWSTF can effectively overcome the degradation of the feature extraction caused by motion blur.

In order to more intuitively evaluate the effectiveness of attachment recognition, a qualitative comparison is made with the SS results of all methods. Figure 4 shows some of the comparison results. As seen in Figure 4, the higher quality can be seen in the

segmentation maps output by the proposed method compared with the other two methods. It is seen that the contour of the segmentation map of IEWSTF is closer to the GT. The other two methods are more obviously affected by motion blur, and distortion is observed in the segmentation maps at various rotational speeds.

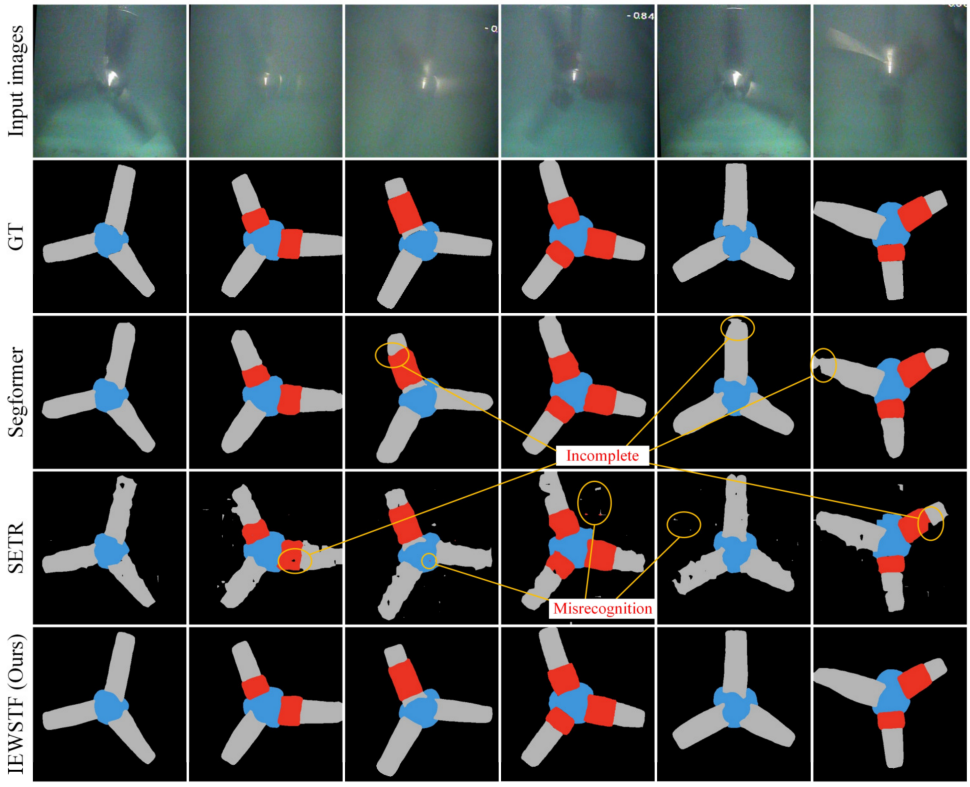

Figure 4. Qualitative comparison of segmentation maps of different methods.

Specifically, incomplete segmentation edges are found in some of the results from SegFormer for the location of the blade tip. This is due to the fact that there is a more significant motion blur at the tip of the blade compared to the root, thus having a greater impact on accuracy. Furthermore, the proposed IEWSTF has complete segmented edges by adaptively learning the ST features between adjacent frames, which can be compensated after being motion-blurred.

For the recognition results of the attachment, both SETR and SegFormer showed incomplete recognition of the coverage area of the attachment, as can be seen from the third result of SegFormer and the second and sixth results of SETR. In contrast, the proposed method has the most accurate recognition of the attachment, with an accurate depiction of the location of the attachment and the boundaries of the coverage area.

In addition, it shows more misrecognitions from the segmentation maps of SETR, especially in the background. This is determined by the mechanisms within SETR. In SETR, the ViT transforms the image into multiple patch embeddings via location. Each patch embedding is a local region in the image. Since these local regions do not overlap each other, they will be more vulnerable to noise due to the lack of learning for common features at the junctions. While IEWSTF adopts the transformation strategy of overlapping patch embedding, which is similar to SegFormer, the common features of the overlapping part of patch embeddings can be learned to eliminate the interference of noise.

4.4. Evaluation of the Robustness against the Variations in Rotational Speed

In order to evaluate the robustness of the proposed method against the variations in rotational speed of the MCT, the models are trained using only a portion of the data of level 1. Then, the models are tested on the remaining level 1 data and all data from level 2

to level 5. Figure 5 shows the comparison of the testing results. As shown in Figure 5, when trained with data from level 1, there is a decrease in the mIoU for all methods from level 2 to level 5. SETR and SegFormer both show declines of more than 30%. Furthermore, while the IEWSTF shows the slightest drop, the mIoU is still above 70% from level 2 to level 5, which indicates that the IEWSTF improves the robustness of the SS network against the variations in rotational speed. It should be noted that when the speed is accelerated, the differences are subsequently increased between the adjacent frames. In this case, the IEW mechanism will adaptively assign smaller weights to the features of the adjacent frames. This also contributes to the higher robustness of the model.

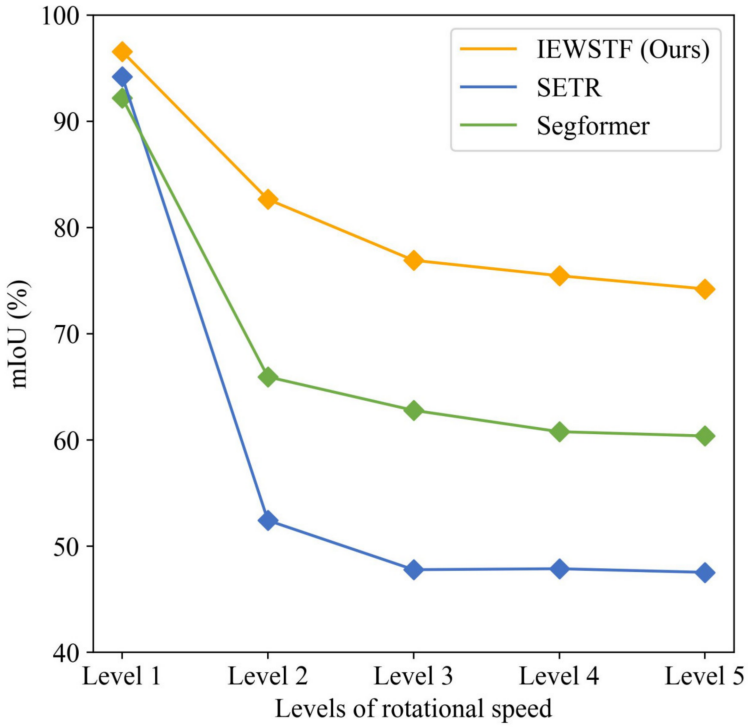

Figure 5. Comparison of mIoU for all methods at different levels of rotational speed.

5. Conclusions

An IEWSTF-based SS method is proposed in this paper to improve the robustness of the SS method against the variations in MCT rotational speeds. The method adaptively learns and fuses the ST features in adjacent frames by using the STF and IEW mechanisms. The method performs excellently with the MCT image datasets with multiple rotational speed levels. Additionally, the proposed method shows stronger robustness to the variations in MCT rotational speeds than the comparison methods. The experimental results also demonstrate that the proposed method can meet the requirements for practical applications with sufficient training data.

Follow-up work can focus on further extending the model to semi-supervised or unsupervised SS methods to reduce the dependence on labeled datasets. In addition, future research could also focus on studying how to reduce the impact of marine environmental variations on the model.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Heidarian, A.; Cheung, S.C.P.; Rosengarten, G. The effect of flow rate and concentration on the electrical conductivity of slurry electrodes using a coupled computational fluid dynamic and discrete element method (CFD–DEM) model. *Electrochem. Commun.* **2021**, *126*, 107017. [CrossRef]
2. Heidarian, A.; Cheung, S.C.P.; Ojha, R.; Rosengarten, G. Effects of current collector shape and configuration on charge percolation and electric conductivity of slurry electrodes for electrochemical systems. *Energy* **2022**, *239*, 122313. [CrossRef]
3. Segura, E.; Morales, R.; Somolinos, J.A. A Strategic Analysis of Tidal Current Energy Conversion Systems in the European Union. *Appl. Energy* **2018**, *212*, 527–551. [CrossRef]
4. Myers, L.; Bahaj, A.S. Simulated electrical power potential harnessed by marine current turbine arrays in the Alderney Race. *Renew. Energy* **2005**, *30*, 1713–1731. [CrossRef]
5. Xie, T.; Wang, T.; He, Q.; Diallo, D.; Claramunt, C. A review of current issues of marine current turbine blade fault detection. *Ocean. Eng.* **2020**, *218*, 108194. [CrossRef]
6. Farkas, A.; Degiuli, N.; Martić, I.; Barbarić, M.; Guzović, Z. The impact of biofilm on marine current turbine performance. *Renew. Energy* **2022**, *190*, 584–595. [CrossRef]
7. Walker, J.S.; Green, R.B.; Gillies, E.A.; Phillips, C. The effect of a barnacle-shaped excrescence on the hydrodynamic performance of a tidal turbine blade section. *Ocean. Eng.* **2020**, *217*, 107849. [CrossRef]
8. Song, S.; Demirel, Y.K.; Atlar, M.; Shi, W. Prediction of the fouling penalty on the tidal turbine performance and development of its mitigation measures. *Appl. Energy* **2020**, *276*, 115498. [CrossRef]
9. Grosvenor, R.I.; Prickett, P.W.; He, J. An Assessment of Structure-Based Sensors in the Condition Monitoring of Tidal Stream Turbines. In Proceedings of the 2017 Twelfth International Conference on Ecological Vehicles and Renewable Energies (EVER), Monte Carlo, Monaco, 11–13 April 2017.
10. Yang, C.; Xie, T.; Wang, T.; Wang, Y.; Jia, D. Blade attachment degree classification for marine current turbines using AGMF-DFA under instantaneous variable current speed. *Proc. Inst. Mech. Eng. Part M J. Eng. Marit. Environ.* **2022**, *in press*. [CrossRef]
11. Zhang, M.; Wang, T.; Tang, T. A multi-mode process monitoring method based on mode-correlation PCA for Marine Current Turbine. In Proceedings of the 2017 IEEE 11th International Symposium on Diagnostics for Electrical Machines, Power Electronics and Drives (SDEMPED), Tinos, Greece, 29 August–1 September 2017.
12. Zhang, M.; Wang, T.; Tang, T.; Benbouzid, M.; Diallo, D. An Imbalance Fault Detection Method Based on Data Normalization and EMD for Marine Current Turbines. *ISA Trans.* **2017**, *68*, 302–312. [CrossRef]
13. Zheng, Y.; Wang, T.; Xin, B.; Xie, T.; Wang, Y. A Sparse Autoencoder and Softmax Regression Based Diagnosis Method for the Attachment on the Blades of Marine Current Turbine. *Sensors* **2019**, *19*, 826. [CrossRef]
14. Xin, B.; Zheng, Y.; Wang, T.; Chen, L.; Wang, Y. A diagnosis method based on depthwise separable convolutional neural network for the attachment on the blade of marine current turbine. *Proc. Inst. Mech. Eng. Part I J. Syst. Control. Eng.* **2021**, *235*, 1916–1926. [CrossRef]
15. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
16. Peng, H.; Yang, D.; Wang, T.; Pandey, S.; Chen, L.; Shi, M.; Diallo, D. An Adaptive Coarse-Fine Semantic Segmentation Method for the Attachment Recognition on Marine Current Turbines. *Comput. Electr. Eng.* **2021**, *93*, 107182. [CrossRef]
17. Zhao, S.; Oh, S.-K.; Kim, J.-Y.; Fu, Z.; Pedrycz, W. Motion-blurred image restoration framework based on parameter estimation and fuzzy radial basis function neural networks. *Pattern Recognit.* **2022**, *132*, 108983. [CrossRef]
18. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
19. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015-Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.
20. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
21. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587v3.
22. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv* **2018**, arXiv:1802.02611.

23. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *arXiv* **2021**, arXiv:2105.15203.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929v2.
25. Li, J.; Chen, J.; Tang, Y.; Wang, C.; Landman, B.A.; Zhou, S.K. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Med. Image Anal.* **2023**, *85*, 102762. [CrossRef]
26. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
27. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
28. Li, W.; Zhao, Y.; Li, F.; Wang, L. MIA-Net: Multi-information aggregation network combining transformers and convolutional feature learning for polyp segmentation. *Knowl.-Based Syst.* **2022**, *247*, 108824. [CrossRef]
29. Wu, R.; Wen, X.; Yuan, L.; Xu, H. DASFTOT: Dual attention spatiotemporal fused transformer for object tracking. *Knowl.-Based Syst.* **2022**, *256*, 109897. [CrossRef]
30. Zhao, Q.; Yang, H.; Zhou, D.; Cao, J. Rethinking Image Deblurring via CNN-Transformer Multiscale Hybrid Architecture. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 1–15. [CrossRef]
31. Hou, Z.; Li, F.; Wang, S.; Dai, N.; Ma, S.; Fan, J. Video object segmentation based on temporal frame context information fusion and feature enhancement. *Appl. Intell.* **2023**, *53*, 6496–6510. [CrossRef]
32. Ji, G.; Fan, D.; Fu, K.; Wu, Z.; Shen, J.; Shao, L. Full-duplex strategy for video object segmentation. *Comput. Vis. Media* **2023**, *9*, 155–175. [CrossRef]
33. Wang, L.; Liu, H.; Zhou, S.; Tang, W.; Hua, G. Instance Motion Tendency Learning for Video Panoptic Segmentation. *IEEE Trans. Image Process.* **2023**, *32*, 764–778. [CrossRef]
34. Zhao, D.; Liu, L.; Yu, F.; Heidari, A.A.; Wang, M.; Liang, G.; Muhammad, K.; Chen, H. Chaotic random spare ant colony optimization for multi-threshold image segmentation of 2D Kapur entropy. *Knowl.-Based Syst.* **2021**, *216*, 106510. [CrossRef]
35. Lei, B.; Fan, J. Adaptive granulation Renyi rough entropy image thresholding method with nested optimization. *Expert Syst. Appl.* **2022**, *203*, 117378. [CrossRef]
36. Naidu, M.S.R.; Rajesh Kumar, P.; Chiranjeevi, K. Shannon and Fuzzy entropy based evolutionary image thresholding for image segmentation. *Alexandria Eng. J.* **2018**, *57*, 1643–1655. [CrossRef]
37. Li, J.; Wang, W.; Chen, J.; Niu, L.; Si, J.; Qian, C.; Zhang, L. Video Semantic Segmentation via Sparse Temporal Transformer. In Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, USA, 20–24 October 2021.
38. Zhang, R.; Xiao, Q.; Du, Y.; Zuo, X. DSPI Filtering Evaluation Method Based on Sobel Operator and Image Entropy. *IEEE Photonics J.* **2021**, *13*, 1–10. [CrossRef]
39. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.O.; Villena-Martinez, V.; Garcia-Rodriguez, J. A review on deep learning techniques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857.