applied sciences

MDPI

Article

Mechanical Assembly Monitoring Method Based on Semi-Supervised Semantic Segmentation

Suichao Wu [1], Chengjun Chen [1,2,*] and Jinlei Wang [1]

[1] School of Mechanical and Automotive Engineering, Qingdao University of Technology, Qingdao 266000, China
[2] Key Lab of Industrial Fluid Energy Conservation and Pollution Control, Ministry of Education, Qingdao University of Technology, Qingdao 266000, China
* Correspondence: chencj@qut.edu.cn

Abstract: Semantic segmentation of assembly images is to recognize the assembled parts and find wrong assembly operations. However, the training of supervised semantic segmentation requires a large amount of labeled data, which is time-consuming and laborious. Moreover, the sizes of mechanical assemblies are not uniform, leading to low segmentation accuracy of small-target objects. This study proposes an adversarial learning network for semi-supervised semantic segmentation of mechanical assembly images (AdvSemiSeg-MA). A fusion method of ASFF multiscale output is proposed, which combines the outputs of different dimensions of ASFF into one output. This fusion method can make full use of the high-level semantic features and low-level fine-grained features, which helps to improve the segmentation accuracy of the model for small targets. Meanwhile, the multibranch structure RFASPP module is proposed, which enlarges the receptive field and ensures the target object is close to the center of the receptive field. The CoordConv module is introduced to allow the convolution to perceive spatial position information, thus enabling the semantic segmentation network to be position-sensitive. In the discriminator network, spectral normalization is introduced. The proposed method obtains state-of-art results on the synthesized assembly depth image dataset and performs well on actual assembly RGB image datasets.

Keywords: assembly monitoring; semantic segmentation; semi-supervised; adversarial learning; multiscale feature fusion

Citation: Wu, S.; Chen, C.; Wang, J. Mechanical Assembly Monitoring Method Based on Semi-Supervised Semantic Segmentation. *Appl. Sci.* **2023**, *13*, 1182. https://doi.org/10.3390/app13021182

Academic Editor: Jin Seo Park

Received: 2 October 2022
Revised: 24 December 2022
Accepted: 9 January 2023
Published: 16 January 2023

1. Introduction

Mechanical assembly is an important step in the processing of mechanical products, and assembly quality directly affects the quality of the product. At present, the manufacturing industry produces large-scale customized products, which makes the assembly process very cumbersome. When operators are faced with extremely complex assembly work, ignoring a small detail may lead to assembly errors and reduce product quality.

Vision-based monitoring is an effective and efficient monitoring method, which has been widely used in industry because of its noncontact and nondestructive detection capability [1]. Cyganek et al. [2] realized the monitoring of the driver's state based on the image of the driver. Negin et al. [3] used a vision-based action recognition method to monitor the behaviors related to autism spectrum disorders (ASDs). Fernández-Robles et al. [4] used images of a micro tool and image processing methods to monitor tool wear. Riego et al. [5] employed an industrial camera to photograph the surfaces of inner and outer surfaces of cylindrical bores and classified the captured surface images and combined them with an integrated learning approach to monitor the wear of milling operations. Additionally, vision-based monitoring can also be used for the assembly monitoring of mechanical parts. As an approach for real-time monitoring of mechanical assembly, assembly monitoring can effectively avoid problems such as assembly errors caused by operator fatigue and other factors and ensure the assembly quality of products.

At present, various methods have been proposed for assembly monitoring. For the finishing of large parts, Kaczmarek et al. [6] proposed a computer vision-based manual assembly monitoring system to monitor the assembly process. Hu et al. [7] developed a monitoring system for human–machine collaborative assembly based on 3D hand pose estimation, and the system was experimentally verified on the assembly lines for power protectors. Chen et al. [8] presented a mechanical assembly monitoring scheme based on change detection, which can monitor the parts with changes in the assembly process from multiple angles and segment them. Riedel et al. [9] designed an assistance system for assembly based on object detection, and experiments were conducted on an explosion-proof tubular lamp. The results indicate that the use of this system effectively reduces assembly errors. Mauricio-Andrés et al. [10] proposed an object-detection-based recognition system for tools, components, and assembly actions to ensure product quality.

Most of the above assembly monitoring methods use object detection. However, when the parts of the assembly are small and dense, it will be difficult to perform object detection on each part. In this case, the method of semantic segmentation can be used instead of object detection for assembly monitoring. Chen et al. [11] have proposed a fully supervised semantic segmentation network for monitoring the mechanical assembly process.

Because the parts of the assembly are dense and of the same color, assemblers are prone to misassemble or missing assembly during the assembly process, especially for small parts. In response to this problem, inspired by the literature [11], this paper uses the semantic segmentation method to identify the pixels of each part of the assembly image and annotates the pixels of different parts with different colors. This facilitates the identification of parts during the assembly process and helps the assembler to check for missing or misplaced parts in the assembly sequence. However, in the training process of semantic segmentation, a large number of accurate per-pixel labeling operations are required. To reduce the number of labels to save costs, this paper applies a semi-supervised method. Compared with the fully supervised method, the semi-supervised method adopts a mixture of a small amount of labeled data and a mass of unlabeled data. However, the size of the parts in the mechanical assembly image is not uniform, and the convolution operation of many downsampling operations in the segmentation network loses the underlying fine-grained information of the image, which leads to low segmentation accuracy.

An adversarial learning network for semi-supervised semantic segmentation of mechanical assembly images (AdvSemiSeg-MA) is proposed in this paper. The AdvSemiSeg-MA network fuses all the scale features of ASFF output into one output, which makes full use of the deep semantic and shallow detail features in the network. Thus, the precision of small object segmentation is improved. Meanwhile, an RFASPP module is proposed, which imitates the multibranch structure of the human receptive field and extracts more deep features. In semantic segmentation, the CoordConv module is introduced to enable convolution to perceive spatial position information and enable the semantic segmentation network to be position-sensitive. In the discriminant network, spectral normalization is introduced, and the discriminant network layer is deepened, which enhances the stability of AdvSemiSeg-MA network training and improves the accuracy. The code and dataset are available at https://github.com/DeeplearningXiaobai/AdvSemiSeg-MA (accessed on 22 September 2022).

2. Related Work

Attributed to the rapid development of deep learning and convolutional neural networks, semantic segmentation technology has developed rapidly. The goal of semantic segmentation is to classify pixels, i.e., semantic segmentation needs to classify each pixel. According to the use of labels, methods of semantic segmentation can be divided into three categories: fully supervised semantic segmentation, unsupervised semantic segmentation, and semi-supervised semantic segmentation. The three categories of methods are briefly introduced below.

2.1. Fully Supervised Semantic Segmentation

Fully supervised learning uses labeled datasets to train models. The fully supervised semantic segmentation method uses accurately labeled datasets for network training. The labeled dataset can enhance detailed information and local features, which helps to enhance the precision of the semantic segmentation network. Long et al. [12] are the first to present a fully convolutional network (FCN) by using convolution instead of full connection to realize image segmentation. FCN directly upsamples the image by a factor of 8, leading to rough segmentation results. For the problem of FCN, Ronneberger et al. [13] proposed a U-Net network, which is a U-shaped structure based on an encoder–decoder. The U-Net network uses multilayer skip connections and upsampling to improve the segmentation effect. To solve the problems of FCN, the DeepLab series takes another approach. In DeepLabv2, to fuse multiscale features, Chen et al. [14] first proposed an atrous spatial pyramid pooling (ASPP) technique. Then, DeepLabv3 [15] and DeepLabv3plus [16] improved the method of ASPP and achieved higher segmentation accuracy. Supervised semantic segmentation demands a mass of labeled data to achieve high semantic segmentation accuracy, but data labeling is very difficult. Therefore, this method is not suitable for industrial applications.

2.2. Unsupervised Semantic Segmentation

Unsupervised learning uses unlabeled datasets to train models. The unsupervised semantic segmentation method can avoid the consumption of human and material resources required by data labeling. Van Gansbeke et al. [17] first performed pixel-level representation learning and then fine-tuned the network to realize unsupervised semantic segmentation. Through a training method using the virtual city dataset as the source domain and the real city dataset as the target domain, Tsai et al. [18] adopted the adversarial idea to achieve unsupervised domain adaptation of the target domain and achieve semantic segmentation of the target domain. In the field of domain adaptation, unsupervised learning is realized by training the network in the source domain and predicting in the target domain. Although the unsupervised semantic segmentation method does not need labeled data, the accuracy of the current unsupervised semantic segmentation method cannot meet the high-precision requirements in the industry.

2.3. Semi-Supervised Semantic Segmentation

Different from supervised learning which needs plenty of labels and unsupervised learning which does not require labels, semi-supervised learning uses a mixture of a few labeled data and a large amount of unlabeled data, and it performs training to make the prediction effect as close as possible to that of fully supervised learning. French et al. [19] proposed to generate false labels by mixing two images to supervise the network. Olsson et al. [20] presented a data enhancement method for semantic segmentation. These methods all perform data augmentation of images, but for assembly images lacking color and texture, they are not useful for assembly image segmentation.

Generative Adversarial Network (GAN) [21] is a commonly used approach to perform semi-supervised semantic segmentation. Hung et al. [22] used adversarial learning to generate pseudo-labels from unlabeled datasets and conducted network training to improve network performance. Hung et al. employed a discriminator of the fully convolutional network to discriminate the input image at the pixel level to generate a confidence image to supervise the training process and realize semi-supervised learning. Mittal et al. [23] adopted a double-branch structure, where the upper branch uses the method of Pauline et al. and the lower branch uses the Mean Teacher classifier, which can effectively reduce the false detection rate and enhance network performance. The main ideology of semi-supervised learning is to train the network through pseudo-labels generated by many unlabeled datasets to improve the segmentation capability of the network further.

Although fully supervised semantic segmentation has high accuracy, it will consume a lot of human and financial resources for data labeling. Unsupervised semantic segmentation does not require data labeling, but the segmentation accuracy is not high, so it cannot meet

the high-precision requirements of assembly monitoring. Inspired by the literature [22], this paper proposes an adversarial learning AdvSemiSeg-MA network, which reduces the number of data annotations while ensuring high accuracy.

3. Overall Framework

In this section, the overall structure of the system is first introduced. Then, three improvement schemes for semantic segmentation are described in detail. Finally, the training process of the network is discussed.

3.1. Structure of the Model

This paper proposes an adversarial learning network for semi-supervised semantic segmentation of mechanical assembly images (AdvSemiSeg-MA) (Figure 1). The semantic segmentation network G consists of the FRASPP deep feature extraction module and the CoordConv spatial perception module. The discriminator network D consists of multiple layers of convolutions using spectral normalization (SN) [24]. Four types of losses are used in network training, namely, the adversarial loss $L_{\mathbf{adv}}$ of the discriminator network D, the spatial cross-entropy loss L_D, the mask cross-entropy loss $L_{\mathbf{semi}}$, and the standard cross-entropy loss $L_{\mathbf{ce}}$ of annotated images. This network uses the semantic segmentation network as the generator of the generative adversarial network. The output class probability map is sent to the discriminator network D. The discriminator network outputs the true and false discriminant values of each pixel to improve the precision of the semantic segmentation network G.

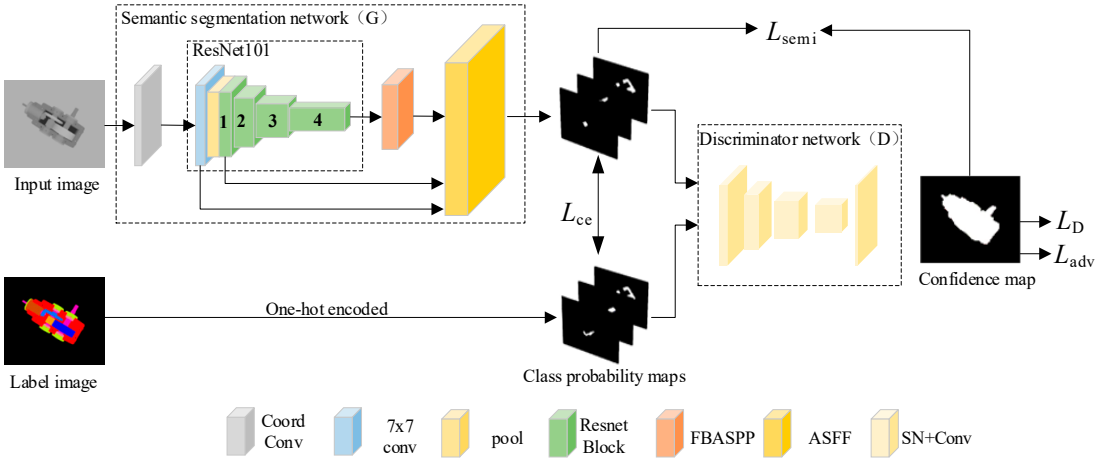

Figure 1. The overall structure of the network.

As shown in Figure 1, a mechanical assembly image X_1 with a size of H × W × 3 is input to G. The image is first input to the CoordConv module so that the convolution can perceive the spatial position information and enable G to be position-sensitive. Then, the obtained features are input to the RseNet101 module to extract the features of the mechanical assembly image. Subsequently, the obtained features are input to the FRASPP module to extract the deep features of the image. Finally, the deep features extracted by the FRASPP module, the features output by the 7 × 7 convolution of RseNet101, and the features output by the first layer of RseNet101 are input to the ASFF module to fuse the above-mentioned multiscale features. A class probability map of dimension H × W × C is the output, denoted as G (X_1), where C is the class count of the semantic segmentation.

In this study, the discriminator network (D) adopts four sets of convolutions including spectral normalization (SN). In addition, a set of convolutions consisting of a 3 × 3 convolution with a step size of 1 and a 4 × 4 convolution with a step size of 2 are added. In D, multilayer spectral normalization (SN) can make D satisfy the Lipschitz continuity. This prevents the function from drastic changes and improves the stability of the training

process. The input of D is the class probability map (G (X_1)) output by G or the one-hot encoded ground truth (Y_n). The output of D is the confidence map of size H × W × 1. For each pixel of the confidence map, it is set to zero if it comes from G and is set to one if it comes from Y_n.

During the semi-supervised training of the AdvSemiSeg-MA network, 1/8 of the assembly image training dataset is randomly used as labeled data, and the remaining data is taken as unlabeled data. Firstly, the G and D are trained using annotated images. The training of G is jointly supervised by the cross-entropy loss L_{ce} of the annotated images and the adversarial loss L_{adv} of D. Meanwhile, the training of D is only supervised by the spatial cross-entropy loss *LD*. Then, the unlabeled images are used to train the network again. The unlabeled images are passed through G to output the class probability map, which is input to D to obtain the confidence map. Finally, the confidence map and mask cross-entropy loss L_{semi} are used as supervised information to train G in a self-learning way.

The innovations of the proposed AdvSemiSeg-MA network include: (1) A multiscale output fusion method of ASFF is proposed. This fusion method can make full use of the deep semantic and shallow detail features extracted by the network to improve the accuracy of the model for segmenting small parts in mechanical assembly. (2) A multibranch RFASPP module that imitates the human receptive field is proposed. It ensures that the target object gets near the center of the receptive field, and the receptive field is enlarged so that the capacity of the model to extract deep features is increased. (3) In G, the CoordConv module is introduced to allow the convolution to obtain spatial position information. This enables G to be position-sensitive. (4) In D, SN is introduced, and the number of convolutional layers of D is deepened. Based on this, the discriminant network satisfies the Lipschitz continuity, thereby improving the stability of network training.

G performs operations such as convolution and pooling on the input assembly image to obtain a class probability map. This study realizes semi-supervised assembly semantic segmentation with G as the generator of the GAN. To obtain a class probability map with higher precision, this study proposes a new semantic segmentation network by adding the ASFF module, FRASPP module, and CoordConv module to RseNet101. In G, the multiscale features output by the ASFF module are fused to fully utilize the features of multiple scales, thereby improving the segmentation accuracy of small parts; the FRASPP module improves the ability to extract deep features; the CoordConv module enables G to be position-sensitive.

3.2. Adaptive Spatial Feature Fusion (ASFF) Module

After the downsampling operation, the objects of different sizes on the images can have a large semantic generation gap, resulting in low segmentation accuracy of small objects. The feature pyramid can generate features of different sizes at different scales. Based on this, various objects can find suitable feature representations at different scales. Meanwhile, the fusion of multiscale features can make full use of the deep semantic and shallow detail features of the network, thus enhancing the segmentation accuracy of the network. Currently, many feature pyramid structures use the FPN [25] method, and concat or element-wise methods are employed to fuse deep semantic and shallow detail features. However, these two fusion methods cannot fully utilize the features of different scales. Therefore, Liu et al. [26] proposed a new feature pyramid method called adaptive spatial feature fusion (ASFF). It multiplies the weights trained by the network and the corresponding features to obtain features of different sizes that fuse multiple scales. ASFF is used in the target detection network YOLOV3, and good results are obtained. The main thought of ASFF is to fuse information of multiple scales. Its main steps include adjusting the features to the same size and calculating the weight map of each feature.

The feature pyramid structure is mostly used in object detection to address the issue of scale change. In the semantic segmentation branch of the Panopic FPN network, He Kaiming [27] fused features of multiple scales and obtained good results. Inspired by this,

this study uses ASFF in AdvSemiSeg-MA semantic segmentation network and fuses the multiscale output of ASFF into one output. As shown in Figure 2, the output of ASFF is upsampled two times, and it is spliced with the features of the corresponding size. Then, 1×1 convolution is used for channel adjustment. The adjusted features are input to the next ASFF module to fuse high-level semantic features and low-level fine-grained features.

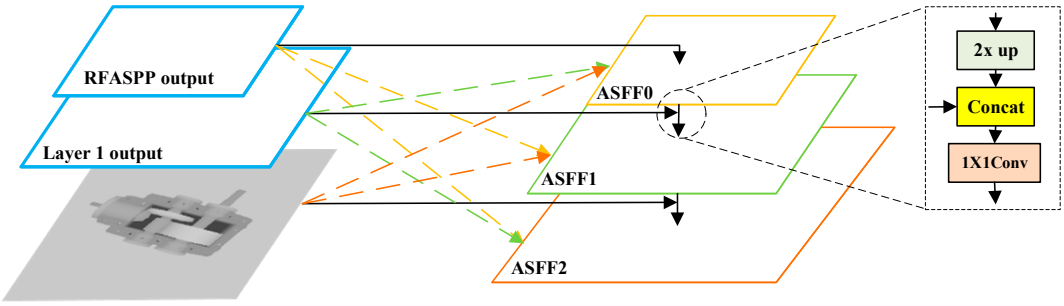

Figure 2. The adaptive spatial feature fusion module.

Specifically, the features obtained after the 7×7 convolution through RestNet101 (7×7 convolution output), the features obtained after the first layer (Layer 1 output), and the features obtained through the FRASPP module (FRASPP output) are input to ASFF0. After upsampling twice, the output of ASFF0 is spliced with the output of the first layer. Meanwhile, a 1×1 convolution is used to adjust the channel. Together with the output of the FRASPP module and the output of the 7×7 convolution, they are input to ASFF1. After upsampling twice, the output of ASFF1 is spliced with the output of 7×7 convolution. In addition, a 1×1 convolution is used to adjust the channel. Together with the Layer 1 output and the FRASPP output, they are input into ASFF2 to obtain a class probability map.

3.3. RFASPP Module

The quality of feature extraction can affect the segmentation performance of the network. RFB [28] simulates the characteristics of receptive fields (RFs) of the human visual system. RFB uses convolutional layers with convolution kernels of different sizes to construct multibranch and dilated convolutional structures of different sizes to simulate the relationship between the dimensions of RFs and the eccentricity rate of RFs. Meanwhile, the reference [28] also points out that different pixels in the RFs have different contributions to the neural nodes, and the center of the RFs contributes the most. Thus, keeping the target object as close as possible to the center of the RFs will enhance the accuracy of the model for segmenting small parts in mechanical assemblies.

In feature extraction, the size of the RFs is important to the extraction ability. When the RF is too small, the local information is overutilized, and the corresponding global information is not obtained. This leads to a correct segmentation of local areas of this category in the image and disorderly segmentation of other areas, thereby reducing the final extraction ability. When the RF is too large, small objects are directly ignored as the background, thereby affecting the segmentation accuracy.

In this study, deeplabv2 is used as the segmentation network of the AdvSemiSeg network, and it adopts the ASPP structure. At the cost of a small amount of calculation, the RF of the convolution kernel is increased, and then more deep features are extracted. To make the target domain close to the center of the RF, this study fuses the RFB and ASPP structures, retains their dominant structures, and proposes a new structure called atrous spatial pyramid pooling based on receptive fields block (RFASPP) (Figure 3). RFASPP removes the shortcut operation in the RFB structure, adds two dilated convolutions with an expansion rate of 18, and adds global pooling. It splices in the channel direction through the concat operation, and finally adjusts the number of channels through 1×1 convolution. The RFASPP structure proposed in this study enlarges the RF while ensuring that the target

object is close to the center of the RF. This can improve the ability of the network to extract deep features.

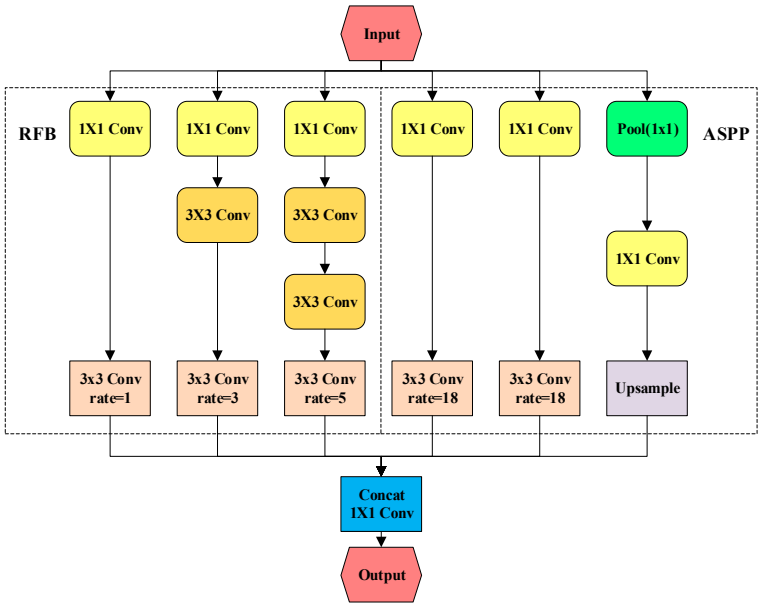

Figure 3. RFASPP module.

3.4. CoordConv Module

Traditional convolution operations are translation invariant. This property improves the robustness of certain tasks such as classification tasks that require spatial invariance. Unlike image classification tasks, semantic segmentation tasks require a position-sensitive convolutional model. The traditional convolution operation, e.g., a local and weight-sharing filter, cannot capture the position information when convolving the feature map. As shown in Figure 4, CoordConv [29] adds two coordinates (*i* and *j*) before convolution to perceive spatial variation information. CoordConv increases the spatial perception ability of convolution by simply adding two channels.

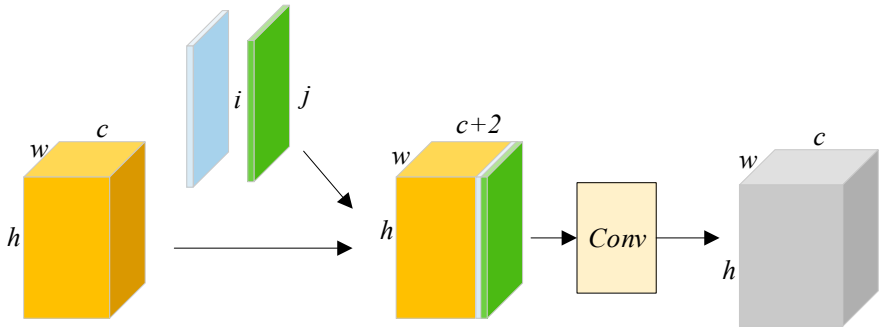

Figure 4. CoordConv module.

G can be regarded as a generator, which generates labels from images. CoordConv can encode high-dimensional concepts such as position in generative models. This helps to improve the performance of generative models [29]. Hence, this study applies CoordConv to the first layer of G to improve the image-to-label generation effect, i.e., to improve the image-to-label segmentation result.

3.5. Network Training

Firstly, to train the discriminator network, it is necessary to minimize the spatial cross-entropy loss L_D. The definition of L_D is as follows:

$$L_D = \sum_{h,w} (1 - y_n) \log(1 - D(G(X_{\mathbf{n}}))^{(h,w)}) + y_n \log(D(Y_{\mathbf{n}})^{(h,w)}), \quad (1)$$

When the input of the discriminator network is the output of the segmentation network, $y_n = 0$; when the input of the discriminator network is the one-hot encoded ground truth, $y_n = 1$. Additionally, $D(\bullet)^{(h,w)}$ refers to the confidence map at location (h,w).

To train the semantic segmentation network, it is necessary to minimize the multitask loss function. The definition of multitask loss is as follows:

$$L_{\mathbf{seg}} = L_{\mathbf{ce}} + \lambda_{\mathbf{adv}} L_{\mathbf{adv}} + \lambda_{\mathbf{semi}} L_{\mathbf{semi}}, \quad (2)$$

where $\lambda_{\mathbf{adv}}$ and $\lambda_{\mathbf{semi}}$ are two weight coefficients and also two hyperparameters for minimizing the loss of multitasking. In this study, the semi-supervised training is divided into two steps: labeled image training and unlabeled image training. Thus, the use of its loss function also has two parts.

When the semantic segmentation network is trained on images with labels, its cross-entropy loss $L_{\mathbf{ce}}$ is defined as follows:

$$L_{\mathbf{ce}} = \sum_{h,w} \sum_{\mathbf{c} \in C} Y_n{}^{(h,w,c)} \log(G(X_{\mathbf{n}})^{(h,w,c)}), \quad (3)$$

where one-hot encoding is applied to convert the discrete real label information mapping into a c-channel probability mapping. Additionally, a fully convolutional discriminator network is used for adversarial learning. Its adversarial loss $L_{\mathbf{adv}}$ is defined as follows:

$$L_{\mathbf{adv}} = \sum_{h,w} \log(D(G(X_{\mathbf{n}}))^{(h,w)}), \quad (4)$$

With the joint supervision of these two losses, the network finishes training on the labeled images.

For unlabeled image training, no labeled images are used for training. Hence, the cross-entropy loss $L_{\mathbf{ce}}$ is no longer used. However, the discriminator network is still needed, so the adversarial loss $L_{\mathbf{adv}}$ is used. Here, the self-learning mechanism is adopted to send the segmentation results of unlabeled data into the discriminator network to generate a confidence map. Meanwhile, a threshold is used to binarize the confidence map, thereby better displaying the area close to the real distribution. Its mask cross-entropy loss $L_{\mathbf{semi}}$ is defined as follows:

$$L_{\mathbf{semi}} = -\sum_{h,w} \sum_{\mathbf{c} \in C} I(D(G(X_{\mathbf{n}}))^{(h,w)} \geq T_{semi} * \overline{Y_n}^{(h,w,c)} \log(G(X_{\mathbf{n}}))^{(h,w,c)}), \quad (5)$$

where $I(\bullet)$ is the indicator function, and T_{semi} is the threshold for self-learning. In addition, if $\mathbf{c}^* = \text{argmax}_{\mathbf{c}} G(X_{\mathbf{n}})^{(h,w,c)}$, the self-learned one-hot encoded labeled image $\overline{Y_{\mathbf{n}}}$ is set element-wise by $\overline{Y_{\mathbf{n}}}^{(h,w,c^*)} = 1$. The self-learn target $\overline{Y_{\mathbf{n}}}$ and the indicator function $I(\bullet)$ are regarded as constants. Experiments show that the network training has good robustness when $T_{semi} = 0.2$.

4. Experiments

4.1. Assembly Image Semantic Segmentation Dataset

In the depth image, each pixel represents the vertical distance between the camera and the object. The depth image is not affected by the interference of environmental changes such as color temperature and illumination, so it is more suitable for industrial environments. Therefore, assembled depth images are used in the study to perform semantic segmentation. Meanwhile, this study takes the assembly process of the secondary bevel gear reducer as an example to verify the effectiveness of the proposed AdvSemiSeg-MA network.

A synthesized assembly depth image dataset is established to verify the effectiveness of the proposed AdvSemiSeg-MA network on depth images. Firstly, SolidWorks is used to establish a 3D model of the reducer, and each part of the assembly model of the secondary bevel gear reducer is saved in the OBJ format. Then, 3Dmax is used to import the assembly model for labeling, and the imaging model of the depth camera and the color camera is established. Finally, the camera's view is changed to shoot to obtain the depth image and label map of the reducer. In this way, the assembly image dataset of the synthesized depth image is established.

As shown in Figure 5, this study divides the assembly process into six stages. Each stage selects 324 images from different views. In the synthesized dataset, there are a total of 1944 depth images and 9 categories. In this study, the dataset is divided into a training set and a test set at a ratio of 9:1. To verify the semi-supervised property of the algorithm proposed in this study, this study randomly selects 1/8 of the data in the training set as labeled data, and the remaining is used as the unlabeled data for training.

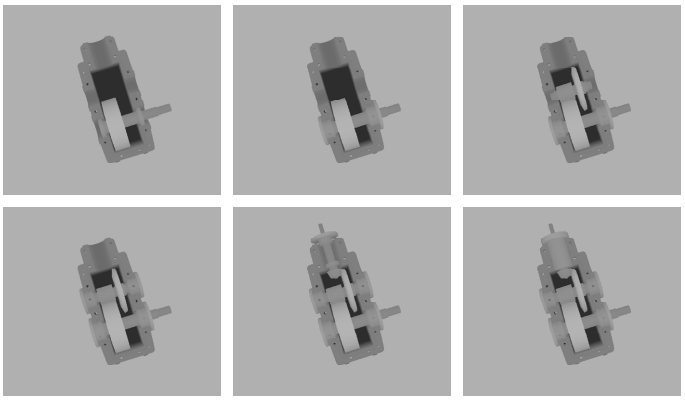

Figure 5. Synthesized assembly depth image dataset.

RGB images are mostly used in the physical environment. Therefore, this paper uses an industrial camera to shoot real RGB images. To obtain an actual assembly RGB image dataset, the assembly was placed on the test bench, and the camera was suspended directly above the assembly to capture images of each assembly process. Then, the images were annotated at the pixel level to obtain the corresponding labels. In this way, an actual assembly RGB image dataset was established. As shown in Figure 6, in each assembly stage, 324 images were selected from different views. In the actual assembly RGB dataset, there are a total of 1620 depth images. In this study, the actual assembly RGB dataset was divided into a training set and a test set at a ratio of 9:1.

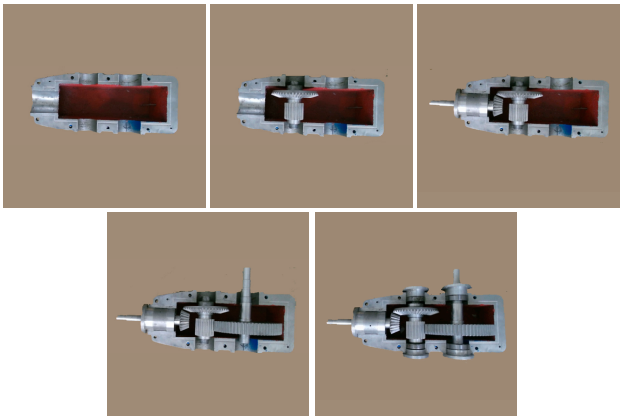

Figure 6. Actual assembly RGB dataset.

The public dataset PASCAL VOC 2012 is also used in this study to further verify the performance of the AdvSemiSeg-MA network. This dataset has a total of 21 categories. A total of 10,582 images of the dataset are included as a training set, and 1449 images are included as a testing set.

In the training process on the synthesized assembly depth image dataset, this study uses depth images and RGB images with a size of 416 × 416 and 512 × 512 for training, and the processing batch of the network is set to 2. In the process of training on the PASCAL VOC 2012 dataset, this study uses images with a size of 321 × 321 for training, and the processing batch of the network is set to 6. In the process of semi-supervised training, 1/8 of the training set is randomly selected as the labeled dataset, and the rest is used as an unlabeled dataset. When training the discriminator network, only the labeled dataset is used for training.

4.2. Evaluation Indicator

There are many evaluation indicators for semantic segmentation. This study selects the most representative PA, MPA, F1, and mean intersection over union (MIoU) as the evaluation indicator of the network. The calculation is as follows:

$$\mathrm{PA} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

$$\mathrm{MPA} = \frac{1}{\mathrm{k}+1}\sum_{\mathrm{i}=0}^{\mathrm{k}} \frac{TP}{TP + FP} \text{ or } \frac{1}{\mathrm{k}+1}\sum_{\mathrm{i}=0}^{\mathrm{k}} \frac{TN}{TN + FN}, \tag{7}$$

$$\mathrm{F1} = \frac{2TP}{2TP + FP + FN}, \tag{8}$$

$$\mathrm{MIoU} = \frac{1}{\mathrm{k}+1}\sum_{\mathrm{i}=0}^{\mathrm{k}} \frac{TP}{TP + FP + FN}, \tag{9}$$

where *FP* is the number of pixels that are incorrectly predicted as positive; *FN* is the number of pixels that are incorrectly predicted as negative; *TP* is the number of pixels that are correctly predicted as positive; *TN* is the number of pixels that are correctly predicted as negative. There are k + 1 classes (0 . . . k) in the dataset, where 0 represents the background.

4.3. Experimental Environment and Parameter Settings

The experimental operating system of this study is Ubuntu 16.04 LTS. The PyTorch [30] framework is used to train the network on a GPU with 8GB video memory. In the segmentation network in the AdvSemiSeg-MA network, the stochastic gradient descent (SGD) optimizer is used. The initial learning rate is set to 2.5×10^{-4}, and its learning rate decreases with a power of 0.9 as a polynomial decay. The network momentum is set to 0.9,

and the weight decay is set to 10^{-4}. For the discriminator network in the AdvSemiSeg-MA network, the Adam optimizer is used. The learning rate is set to 10^{-4}. Its polynomial decay is the same as that of the segmentation network. For hyperparameters, when training on labeled data and unlabeled data, λ_{adv}, respectively, is set to 0.01 and 0.001, λ_{semi} is set to 0.1, and Tsemi is set to 0.2. The AdvSemiSeg-MA network is trained for 20,000 iterations, and the COCO dataset is used to pretrain the weights.

4.4. Ablation Experiments

Ablation experiments are conducted to verify the role of each module in the proposed AdvSemiSeg-MA network. Combined with four innovative works, the performance of each module is verified on the assembly image dataset established in this study, and semi-supervised training is adopted. In the synthesized assembly depth image dataset, 1/8 of the training set is randomly selected as the labeled dataset, and the rest is used as an unlabeled dataset. In addition, this study focuses on the segmentation results of the assembly. Hence, background pixels are removed when calculating MIoU. The ablation experiments use the assembly image dataset to train the network, and there are five experiments in total: Experiment 1 (T1) uses the AdvSemiSeg network. Experiment 2 (T2) uses the AdvSemiSeg + ASFF network. Experiment 3 (T3) uses the AdvSemiSeg + ASFF + RFASPP network. Experiment 4 (T4) uses the AdvSemiSeg + ASFF + RFASPP + CoordConv network. Experiment 5 (T5) uses the AdvSemiSeg-MA network proposed in this study, i.e., based on Experiment 4, the convolution of the discriminator network is spectrally normalized and the depth is deepened.

The results of each experiment on the assembly image dataset are presented in Tables 1 and 2. In this study, the AdvSemiSeg network in Experiment 1 (T1) is used as the baseline. The effect image of the various experiments on the assembly image datasets is shown in Figure 7.

Table 1. The results of each experiment on the synthesized assembly depth image dataset.

Methods	PA/%	MPA/%	F1/%	MIoU/%	Time/s
T1	99.465	96.936	98.108	93.245	0.087
T2	99.837	98.473	99.307	97.000	0.109
T3	99.841	98.841	99.318	97.348	0.113
T4	99.870	98.937	99.341	97.525	0.180
T5 (Ours)	99.886	99.043	99.401	97.728	0.181

Table 2. The results of each experiment on the actual assembly RGB image dataset.

Methods	PA/%	MPA/%	F1/%	MIoU/%	Time/s
T1	99.438	94.231	97.619	89.676	0.150
T2	99.672	97.883	98.580	95.081	0.188
T3	99.674	97.956	98.674	95.199	0.199
T4	99.668	97.986	98.703	95.370	0.238
T5 (Ours)	99.684	97.808	98.883	95.412	0.237

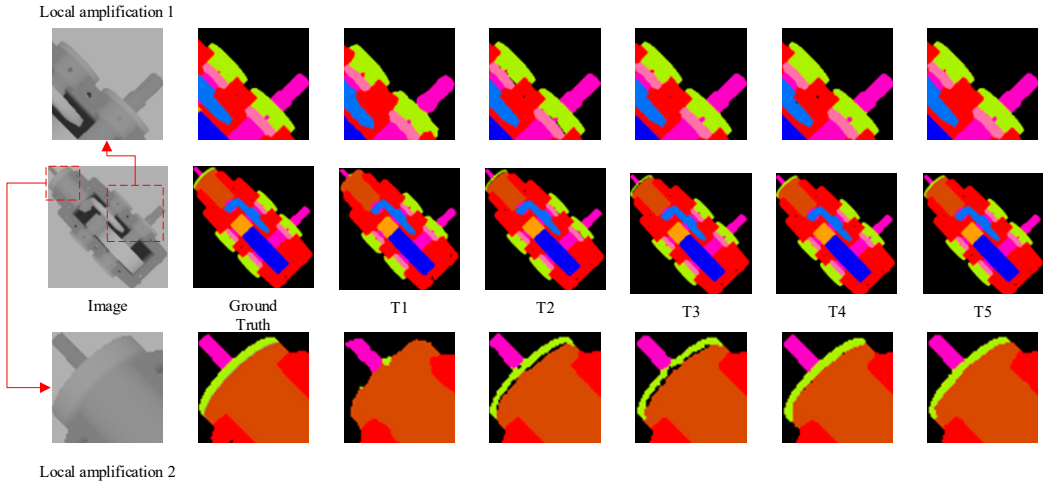

Figure 7. The effect images and local amplified images in each experiment.

As shown in Table 1, by comparing the results of Experiments T1 and T2, it can be seen that after adding the ASFF module, compared with the T1 network, the MIoU of the T2 network increases by 3.755%, the PA increases by 0.372%, the MPA increases by 1.537%, and the F1 increases by 1.199%. In Figure 7, comparing the segmentation results of Experiments T1 and T2, it can be observed that the application of the ASFF multiscale output fusion method improves the boundary segmentation accuracy of the image. This indicates that the ASFF fusion method can enable the network to make full use of the features of different scales and effectively fuse high-dimensional and low-dimensional features.

Meanwhile, the comparison of the results of Experiments T2 and T3 indicates that after adding the RFASPP module, compared with the T2 network, the MIoU of the T3 network increases by 0.348%, the PA increases by 0.004%, the MPA increases by 0.368%, and the F1 increases by 0.011%. In Figure 7, comparing the local amplified image 1 in the segmentation results of T2 and T3, it can be observed that adding the RFASPP module can reduce the fault. This demonstrates that the RFASPP module can improve the ability of the model to extract deep features.

Moreover, comparing the results of Experiments T3 and T4, it can be seen that after adding the CoordConv module, compared with that of the T3 network, the MIoU of the T4 network increases by 0.177%, the PA increases by 0.029%, the MPA increases by 0.096%, and the F1 increases by 0.023%. In Figure 7, comparing the local amplified images 1 and 2 in the segmentation results of Experiments T3 and T4, it can be observed that the addition of the CoordConv module can further reduce the fault. This indicates that the CoordConv module adding the position into the convolution makes the convolution have spatial position perception ability and enables the semantic segmentation network to be position-sensitive.

Moreover, by comparing the results of Experiments T4 and T5, it can be seen that after adding SN to the discriminator network, compared with the T4 network, the MIoU of the T5 network increases by 0.203%, the PA increases by 0.016%, the MPA increases by 0.106%, and the F1 increases by 0.06%. In Figure 7, comparing the local amplified images 1 and 2 in the segmentation results of Experiments T4 and T5, it can be observed that deepening the discriminator network and adding spectral normalization can alleviate the problem that a single pixel is predicted as the background. In addition, these optimizations make the discriminator network satisfy the Lipschitz continuity and improve the training stability of the generative adversarial network.

The time in Table 1 is the average reasoning time for each method to segment one image. It can be seen that the addition of the corresponding modules will increase reasoning time. However, the reasoning time of each image by various methods still meets the needs

of product assembly monitoring. In addition, the accuracy of image segmentation is greatly improved after the introduction of the corresponding modules.

In Table 2, on the actual assembly RGB image dataset, the improved method in this paper can also gradually improve the segmentation accuracy. Compared with the baseline AdvSemiSeg, the MIoU of AdvSemiSeg-MA increases by 5.736%, the PA increases by 0.246%, the MPA increases by 3.577%, and the F1 increases by 1.264%. The AdvSemiSeg-MA network improves the segmentation accuracy of small-target objects in mechanical assembly.

In order to explore the effect of each module on AdvSemiSeg, we conduct experiments by adding each module into AdvSemiSeg. The experimental results can be found in Table 3. As shown in Table 3, all modules except the ASFF module cause degradation in raw network performance. However, adding the modules gradually improves the performance of the model. The experimental results are analyzed as follows. First, the RFASPP module focuses more on extracting high-level semantic features, without the low-dimensional information fusion of the ASFF module, and it loses low-level fine-grained features in low dimensions, so it causes the model performance to be inferior to that of the original network. Secondly, the impact of CoordConv on network performance is related to the location in the network structure, and the location of the CoordConv module is different for different network structures. Though this paper finds the best location to place the CoordConv module for the AdvSemiSeg-MA model by experiment, it is not applicable to the AdvSemiSeg model. Finally, spectral normalization is used to improve the discriminant accuracy of the discriminator. The discriminator is well trained, and the gradient of the generator disappears severely, causing the segmentation performance to decrease

Table 3. The results of each module on the synthesized assembly depth image dataset.

Methods	PA/%	MPA/%	F1/%	MIoU/%
AdvSemiSeg	99.465	96.936	98.108	93.245
AdvSemiSeg + ASFF	99.837	98.473	99.307	97.000
AdvSemiSeg + RFASPP	99.264	95.178	97.450	91.388
AdvSemiSeg + CoordConv	99.160	94.500	96.957	90.560
AdvSemiSeg + spectrally normalized	99.259	93.913	97.684	92.503

4.5. Comparison Experiments on the Assembly Image Dataset

To verify the validity of the AdvSemiSeg-MA network proposed in this study, it is compared with AdvSemiSeg, S4GAN [23], ClassMix [20], PS-MT [31], and U2PL [32]. Among them, PS-MT and U2PL networks are recently proposed and achieve good accuracy in semi-supervised semantic segmentation. Because the RGB images of physical assemblies are manually labeled, there are some errors in the labeling images. Therefore, this study uses the synthesized assembly depth image dataset as the training dataset in comparative experiments, and the synthesized assembly depth image dataset established in this study is used as the training set. The accuracy of the comparison experiments is presented in Table 4.

Table 4. Comparison of semi-supervised methods and our method for segmentation based on the synthesized assembly depth image dataset.

Network	Batch Size	Iteration Times	Pretraining Weight	PA/%	F1/%	MIoU/% (Removing Background)
AdvSemiSeg	2	20,000	COCO	99.465	98.108	93.245
S4GAN	2	40,000	COCO	99.373	98.181	92.918
ClassMix	2	40,000	COCO	99.192	96.945	90.968
PS-MT	2	61,776	COCO	99.504	98.076	93.523
U2PL	2	78,560	COCO	99.659	98.440	94.123
AdvSemiSeg-MA (Ours)	2	20,000	COCO	99.886	99.401	97.728

In Table 4, the number of iterations for training is the original parameter of each network. Meanwhile, the same batch size of 2 is set, and the COCO dataset is used to pretrain the weights. The remaining parameters of each network are set to the optimal parameters of each network. The background is removed when calculating MIoU, and 1/8 of the training set is randomly selected as the labeled data when training the network. Compared with the MIoU values of each network, the AdvSemiSeg-MA network proposed in this study achieves the best segmentation effect on the synthesized assembly depth image dataset. As shown in Table 4, the experimental results indicate that the MIoU, PA, and F1 of the proposed AdvSemiSeg-MA network on the synthesized assembly depth image dataset reaches 97.728%, 99.886%, and 99.401%, which is 3.605%, 0.227%, and 0.961% higher than the current STOA semi-supervised semantic segmentation network (U2PL), respectively. On the synthesized assembly depth image dataset, the approach used in this paper outperforms the current SOTA semi-supervised semantic segmentation network. This validates the effectiveness of the methods adopted by the proposed AdvSemiSeg-MA network, including the multiscale feature fusion method, the dilated convolution pyramid module RFASPP that imitates the human receptive field, the CoordConv module that allows convolution to obtain position information, and the spectral normalization method that improves the training stability. These modules and methods help to improve the segmentation accuracy on the assembly image dataset.

4.6. Comparison Experiments on Public Datasets

This study selects the public dataset PASCAL VOC 2012 to further verify the performance of the AdvSemiSeg-MA network. For fair competition, only the impact of the network structure on the segmentation accuracy is considered. Thus, the same network parameters, the same training process, and the same validation datasets are used. Meanwhile, training and verification are performed on the same computer. The experimental results are shown in Table 5, and this study takes the experimental results of the AdvSemiSeg network in references [23] and [33] for comparison. As shown in Table 5, under the condition of 6 batches and 20,000 iterations, compared with the baseline AdvSemiSeg, the MIoU of AdvSemiSeg-MA increases by 1.879%, the PA increases by 0.523%, and the F1 increases by 0.936%. In addition, in the case of a small number of batches and iterations, the AdvSemiSeg-MA network proposed in this study can still maintain a high segmentation accuracy. Figure 8 shows the comparison of the original images, the labeled images, the segmentation result images obtained by training the AdvSemiSeg network, and the segmentation result images of the method proposed in this study. It can be seen that the proposed AdvSemiSeg-MA network is more effective than the AdvSemiSeg method for the segmentation of image details. Therefore, the AdvSemiSeg-MA network proposed in this study has certain generality.

Figure 8. The experimental effect images of the VOC dataset.

Table 5. Comparison of AdvSemiSeg methods and our method for segmentation on the VOC 2012 dataset.

Network	Batch Size	Iteration Times	PA/%	F1/%	MIoU/%
AdvSemiSeg	6	20,000	92.889	88.288	68.149
AdvSemiSeg [33]	14	-	-	-	69.5
AdvSemiSeg [23]	8	35,000	-	-	69.5
AdvSemiSeg-MA (Ours)	6	20,000	93.412	89.224	70.027

5. Conclusions

This study proposes an adversarial learning network called AdvSemiSeg-MA for semi-supervised semantic segmentation of mechanical assembly images. First, a fusion method of ASFF multiscale output is proposed. This fusion method enables the network to make full use of feature information at different scales and improves the segmentation accuracy of the network for small-target objects. Then, an RFASPP module of the dilated convolutional pyramid that imitates the human receptive field is proposed to improve the network's ability to extract deep features. Subsequently, in the semantic segmentation network, the CoordConv module is introduced to enable the convolution to have spatial perception ability and make the semantic segmentation network position-sensitive. Finally, in the discriminator network, the method of spectral normalization is introduced into the discriminator network, and the depth of the discriminant network is deepened. This improves the stability of semi-supervised network training and enhances the segmentation accuracy. This study establishes the synthesized assembly depth image dataset and actual assembly RGB image dataset for semantic segmentation. The experimental results show that the MIoU of the AdvSemiSeg-MA network on the synthesized assembly depth image dataset and actual assembly RGB image dataset are 97.728% and 95.412%, which are 4.483% and 5.736% higher than that of the baseline AdvSemiSeg, respectively. To further validate the segmentation accuracy, the AdvSemiSeg-MA network is compared with the recently proposed PS-MT and U2PL networks on the synthesized assembly depth image dataset. The results show that the MIoU of AdvSemiSeg-MA is 4.205% higher than that of the PS-MT network and 3.605% higher than that of the U2PL network. Meanwhile, to verify the effectiveness of the network, it is tested on a public dataset. AdvSemiSeg-MA is 1.879% higher than the baseline AdvSemiSeg. Therefore, the AdvSemiSeg-MA network proposed in this study can achieve high segmentation accuracy even when there are only a few labeled datasets and the segmentation target is small.

Author Contributions: Conception of project, C.C. and S.W.; execution, C.C. and S.W.; manuscript preparation, S.W. and J.W.; writing of the first draft, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: The research is supported by the National Natural Science Foundation of China: 52175471.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this research are available on GitHub at https://github.com/DeeplearningXiaobai/AdvSemiSeg-MA (accessed on 22 September 2022).

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

References

1. Shirmohammadi, S.; Ferrero, A. Camera as the Instrument: The Rising Trend of Vision Based Measurement. *IEEE Instrum. Meas. Mag.* **2014**, *17*, 41–47. [CrossRef]
2. Cyganek, B.; Gruszczyński, S. Hybrid Computer Vision System for Drivers' Eye Recognition and Fatigue Monitoring. *Neurocomputing* **2014**, *126*, 78–94. [CrossRef]
3. Negin, F.; Ozyer, B.; Agahian, S.; Kacdioglu, S.; Ozyer, G.T. Vision-Assisted Recognition of Stereotype Behaviors for Early Diagnosis of Autism Spectrum Disorders. *Neurocomputing* **2021**, *446*, 145–155. [CrossRef]

4. Fernández-Robles, L.; Sánchez-González, L.; Díez-González, J.; Castejón-Limas, M.; Pérez, H. Use of Image Processing to Monitor Tool Wear in Micro Milling. *Neurocomputing* **2021**, *452*, 333–340. [CrossRef]
5. Riego, V.; Castejon-Limas, M.; Sanchez-Gonzalez, L.; Fernandez-Robles, L.; Perez, H.; Diez-Gonzalez, J.; Guerrero-Higueras, Á.-M. Strong Classification System for Wear Identification on Milling Processes Using Computer Vision and Ensemble Learning. *Neurocomputing* **2021**, *456*, 678–684. [CrossRef]
6. Kaczmarek, S.; Hogreve, S.; Tracht, K. Progress Monitoring and Gesture Control in Manual Assembly Systems Using 3D-Image Sensors. *Procedia CIRP* **2015**, *37*, 1–6. [CrossRef]
7. Hu, J.-J.; Li, H.-C.; Wang, H.-W.; Hu, J.-S. 3D Hand Posture Estimation and Task Semantic Monitoring Technique for Human-Robot Collaboration. In Proceedings of the 2013 IEEE International Conference on Mechatronics and Automation, Kagawa, Japan, 4–7 August 2013; pp. 797–804.
8. Chen, C.; Li, C.; Li, D.; Zhao, Z.; Hong, J. Mechanical Assembly Monitoring Method Based on Depth Image Multiview Change Detection. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5013413. [CrossRef]
9. Riedel, A.; Gerlach, J.; Dietsch, M.; Herbst, S.; Engelmann, F.; Brehm, N.; Pfeifroth, T. A Deep Learning-Based Worker Assistance System for Error Prevention: Case Study in a Real-World Manual Assembly. *Adv. Prod. Eng. Manag.* **2021**, *16*, 393–404. [CrossRef]
10. Zamora-Hernández, M.-A.; Castro-Vargas, J.A.; Azorin-Lopez, J.; Garcia-Rodriguez, J. Deep Learning-Based Visual Control Assistant for Assembly in Industry 4.0. *Comput. Ind.* **2021**, *131*, 103485. [CrossRef]
11. Chen, C.; Zhang, C.; Wang, T.; Li, D.; Guo, Y.; Zhao, Z.; Hong, J. Monitoring of Assembly Process Using Deep Learning Technology. *Sensors* **2020**, *20*, 4208. [CrossRef] [PubMed]
12. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
13. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
14. 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.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
15. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
16. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
17. Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Van Gool, L. Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10052–10062.
18. Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; Chandraker, M. Learning to Adapt Structured Output Space for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7472–7481.
19. French, G.; Laine, S.; Aila, T.; Mackiewicz, M.; Finlayson, G. Semi-Supervised Semantic Segmentation Needs Strong, Varied Perturbations. *arXiv* **2019**, arXiv:1906.01916.
20. Olsson, V.; Tranheden, W.; Pinto, J.; Svensson, L. Classmix: Segmentation-Based Data Augmentation for Semi-Supervised Learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2021; pp. 1369–1378.
21. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
22. Hung, W.-C.; Tsai, Y.-H.; Liou, Y.-T.; Lin, Y.-Y.; Yang, M.-H. Adversarial Learning for Semi-Supervised Semantic Segmentation. *arXiv* **2018**, arXiv:1802.07934.
23. Mittal, S.; Tatarchenko, M.; Brox, T. Semi-Supervised Semantic Segmentation with High-and Low-Level Consistency. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1369–1379. [CrossRef] [PubMed]
24. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral Normalization for Generative Adversarial Networks. *arXiv* **2018**, arXiv:1802.05957.
25. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
26. Liu, S.; Huang, D.; Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. *arXiv* **2019**, arXiv:1911.09516.
27. Kirillov, A.; Girshick, R.; He, K.; Dollár, P. Panoptic Feature Pyramid Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 6399–6408.
28. Liu, S.; Huang, D. Receptive Field Block Net for Accurate and Fast Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
29. Liu, R.; Lehman, J.; Molino, P.; Petroski Such, F.; Frank, E.; Sergeev, A.; Yosinski, J. An Intriguing Failing of Convolutional Neural Networks and the Coordconv Solution. *arXiv* **2018**, arXiv:1807.03247v2.

30. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
31. Liu, Y.; Tian, Y.; Chen, Y.; Liu, F.; Belagiannis, V.; Carneiro, G. Perturbed and Strict Mean Teachers for Semi-Supervised Semantic Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4258–4267.
32. Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; Le, X. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4248–4257.
33. Alonso, I.; Sabater, A.; Ferstl, D.; Montesano, L.; Murillo, A.C. Semi-Supervised Semantic Segmentation with Pixel-Level Contrastive Learning from a Class-Wise Memory Bank. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 8219–8228.