

Article

Swin-MFA: A Multi-Modal Fusion Attention Network Based on Swin-Transformer for Low-Light Image Human Segmentation

Xunpeng Yi , Haonan Zhang, Yibo Wang , Shujiang Guo, Jingyi Wu and Cien Fan ^{*}

Electronic Information School, Wuhan University, Wuhan 430064, China

* Correspondence: fce@whu.edu.cn

Abstract: In recent years, image segmentation based on deep learning has been widely used in medical imaging, automatic driving, monitoring and security. In the fields of monitoring and security, the specific location of a person is detected by image segmentation, and it is segmented from the background to analyze the specific actions of the person. However, in low-illumination conditions, it is a great challenge to the traditional image-segmentation algorithms. Unfortunately, a scene with low light or even no light at night is often encountered in monitoring and security. Given this background, this paper proposes a multi-modal fusion network based on the encoder and decoder structure. The encoder, which contains a two-branch swin-transformer backbone instead of the traditional convolutional neural network, fuses the RGB and depth features with a multiscale fusion attention block. The decoder is also made up of the swin-transformer backbone and is finally connected via the encoder with several residual connections, which are proven to be beneficial in improving the accuracy of the network. Furthermore, this paper first proposes the low light-human segmentation (LLHS) dataset of portrait segmentation, with aligned depth and RGB images with fine annotation under low illuminance, by combining the traditional monocular camera and a depth camera with active structured light. The network is also tested in different levels of illumination. Experimental results show that the proposed network has good robustness in the scene of human segmentation in a low-light environment with varying illumination. The mean Intersection over Union (mIoU), which is often used to evaluate the performance of image segmentation model, of the Swin-MFA in the LLHS dataset is 81.0, is better than those of ACNet, 3DGNN, ESANet, RedNet and RFNet at the same level of depth in a mixed multi-modal network and is far ahead of the segmentation algorithm that only uses RGB features, so it has important practical significance.



Citation: Yi, X.; Zhang, H.; Wang, Y.; Guo, S.; Wu, J.; Fan, C. Swin-MFA: A Multi-Modal Fusion Attention Network Based on Swin-Transformer for Low-Light Image Human Segmentation. *Sensors* **2022**, *22*, 6229. <https://doi.org/10.3390/s22166229>

Academic Editors: Miguel García-Pineda and Miguel Arevalillo-Herráez

Received: 30 July 2022

Accepted: 17 August 2022

Published: 19 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: multi-modal fusion network; segmentation; low light environment; depth-sensing

1. Introduction

Image segmentation is an important subject in the field of computer vision, the purpose of which is to segment specific objects from various backgrounds [1,2]. The edge, color, texture and many other features of the image are used as the standard to segment the image into different regions by the traditional methods. For example, based on threshold [3], edge [4] and clustering [5,6], these traditional segmentation methods are relatively simple but cannot segment images accurately in complex scenes. Therefore, image segmentation based on deep learning with a higher accuracy has become a research hotspot.

Starting from the proposal of a fully convolutional neural network (FCN) [7], semantic segmentation algorithms based on neural networks have appeared on the stage. An FCN extends the image classification task to the image segmentation task with the pixel level, which lays a foundation for the current image semantic segmentation research. At present, in order to enhance the effect of semantic segmentation, there is more and more research on new semantic segmentation based on neural networks.

With the improvement of the efficiency and accuracy of image segmentation, image segmentation, especially human segmentation, has shown a wide application prospect in

many fields. However, in order to ensure safety and reliability in the fields for automobile navigation and security monitoring, image segmentation is often required to have the ability to process the images of various scenes during the whole day and even at night. In a night scene, the RGB image often has uneven illumination, low light intensity or even no light at all. An image segmentation algorithm only based on the RGB feature usually cannot work well in this situation. Therefore, many recent studies have also carried out attempts to solve this problem, including using mixed datasets of day and night for adversarial training [8] and the introduction of thermal images for multi-modal fusion [9]. However, these studies are all passive imaging, still subject to environmental constraints. The RGB-D cameras that have been developed in recent years can collect the depth data of scenes through active structured light, which means they are less dependent on the environment itself, providing a new possibility for night-time image segmentation.

At present, research on RGB-D image segmentation mainly focuses on solving the fusion problem of the RGB image and depth image and the inaccurate measurement problem of depth images [10], while little research notices its potential application in low-illumination scenes. Therefore, this paper conducts research on image segmentation at night and proposes a transformer-based neural network and feature-fusion attention mechanism. The self-attention mechanism is used to replace a traditional convolutional neural network (CNN), with the purpose of realizing compensation for the loss of information in RGB images with depth information and, finally, achieve the goal of image segmentation in low-illumination scenes. Notice that the existing RGB-D datasets, such as SUN RGB-D, Cityscapes and NYU Depth V2, are not dedicated datasets for low-illuminance conditions as shown in Figure 1. In addition, unfortunately, the existing low-light datasets have either a single background or poor quality. This paper introduces a human segmentation dataset in low light scenes.

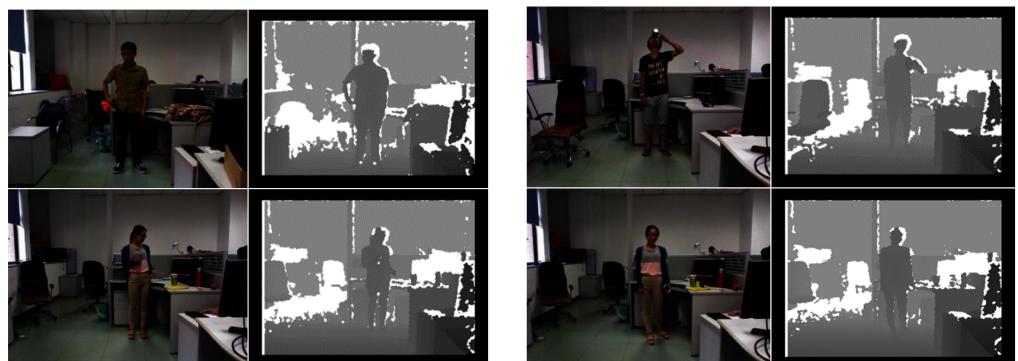


Figure 1. The existing RGB-D datasets focused on human segmentation.

2. Relative Work

At present, single-mode image segmentation has made great progress in segmentation accuracy and efficiency. In ref. [11], U-net uses the short channels to splice encoding and decoding parts, which retain more original information, solving the problem of gradient disappearance to some extent. In ref. [12], ICnet, which uses multi-scale images as input and a cascaded network to improve efficiency, was proposed. Moreover, ICnet limits the input size of images by scaling, thus increasing the reasoning speed of the network.

A transformer was initially applied in natural language processing (NLP) [13]. Recently, many works have transplanted the transformer to CV tasks and achieved good results [14]. In ref. [15], a Vision Transformer was proposed, which cut the serialized data of images into small pieces as the input of the transformer, demonstrating the powerful capability of a transformer in the field of computer vision. In ref. [16], a Vision Transformer using shifted window was proposed, which has strong performance in image classification. Moreover, a Dense-Transformer was proposed to capture the sequence spectral relationship in ref. [17], realizing hyperspectral image (HSI) classification. Some reinforced transform-

ers, such as RTN [18], were used for the automatic quality evaluation of medical images. Transformers were used in the image segmentation in the earlier period in ref. [19], where the transformer was used to completely replace the encoding part of a traditional FCN with the attention mechanism. However, its application in multi-mode and multi-feature was still lacking. Moreover, the transformer's performance in variable lighting and low-lighting conditions remains to be seen.

At present, there has been much research on RGB-D image segmentation, but little research on RGB-D complementarity in low-illumination conditions. Moreover, some RGB-D datasets are of low quality and without fine annotation. In ref. [20], NIN network was used to extract depth image features and integrate them into the GoogLeNet network. In ref. [21], the LDFNet was proposed, which incorporates luminance, depth and color information by a fusion network. In ref. [22], a 3D graph neural network (3DGNN) was proposed to construct a k-nearest neighbor graph based on a KNN pair 3D-point-cloud graph. In ref. [23], three ResNet network branches are adopted, in which two branches are used to extract RGB and depth image features, and the other branch is used to fuse RGB and depth image features. However, the relevant performance of network image segmentation at night and other complex scenes has not been studied, which needs further research and confirmation.

In short, in order to solve the problem that the existing image-segmentation methods cannot be applied to low illumination scenes, the contributions of this paper are as follows:

- Human body images are segmented in the multi-modal and multi-feature way in low-illumination scenes, by using the fusion information of the depth image and RGB image as the segmentation basis.
- A multi-modal end-to-end segmentation network based on swin-transformer is proposed, which realizes end-to-end RGB and depth feature-fusion attention by combining swin-transformer features that are demonstrated to be stable under changeable-lighting conditions. It can totally replace the traditional convolutional neural network and improve the accuracy of segmentation.
- Aiming at the shortcomings of traditional image segmentation under low illumination, a modified and pre-processed body semantic segmentation dataset (LLHS) with fine annotation for a low-light scene is proposed, which is much larger in scale and scene than the previous dataset, filling the gap in the semantic segmentation dataset in the low-illuminance condition.

3. Materials and Methods

Swin-MFA proposed in this paper is an end-to-end multi-modal segmentation network with low illumination optimization. This network adopts encoder and decoder structure with transformer backbone, which has good noise tolerance and accuracy for human segmentation in low light conditions. In addition, the dataset LLHS was produced for the deep multi-modal method to solve the problem of human body segmentation in low-illumination scenes, which has good advantages in terms of the size and quality.

3.1. Low Light Human Segmentation Dataset

Low light–human segmentation dataset is a new portrait dataset in low-light scenes, which adopts active ranging sensing method to collect depth images based on structured light principle and collect RGB images by a traditional RGB camera. Due to the black vacancy at the edge of the portrait in the depth image and the registration problem between the depth image and RGB image, the dataset is preprocessed as follows.

- (1) The physical location of the camera of RGB image and depth image results in different spatial-coordinate systems. The images taken by RGB camera and depth camera are not matched by pixels, so it is necessary to register RGB image and depth image. The internal parameter matrix and external parameter matrix in different scenes are obtained by calibrating RGB camera and depth camera, respectively. Then, the transformation matrix of two coordinate systems is calculated by Equation (1):

$$p_{rgb} = H_{rgb}(RH_{ir}^{-1}p_{ir} + T) \quad (1)$$

where, p_{ir} is the coordinates of pixels in the depth image before processing, H_{ir} is the internal parameter matrix of the depth camera and H_{rgb} is the internal parameter matrix of the RGB camera. R and T are rotation matrices and shift vectors, respectively, derived from the outer parameter matrix.

$$R = R_{rgb}R_{ir}^{-1} \quad (2)$$

$$T = T_{rgb} - R_{rgb}R_{ir}^{-1}T_{ir} \quad (3)$$

where $R_{ir}(R_{rgb})$ and $T_{ir}(T_{rgb})$ are rotation matrix and shift vector of depth camera (RGB camera) in external parameter matrix, respectively.

- (2) In depth images, due to camera shooting angle and objects blocking, black gaps appear in the image, resulting in the interference of the image edge information, which needs to be processed. The depth camera of Realsense device is set on the left side, and the imaging algorithm is realized by referring to the left camera. Therefore, the upper and lower five pixels of the left side adjacent to the black gap can be used as the processing neighborhood to fill the vacancy. In order to maintain image-edge information, it is necessary to make the filled pixels contain background information rather than foreground information. Therefore, the pixel of the farthest point with the largest pixel value in the neighborhood is used to fill the black vacancy. The specific calculation formula can be expressed in Equation (4).

$$P'_{(i,j)} = \text{MAX}\left\{P_{(i-1,j)}, P_{(i+1,j)}, P_{(i,j-1)}, P_{(i-1,j-1)}, P_{(i-1,j+1)}\right\} \quad (4)$$

where $P_{(i,j)}$ is the pixel value of the i -th row and j -th column in the filling kernel, and $P'_{(i,j)}$ is the corresponding pixel value of the i -th row and j -th column in the image after processing.

The corrected RGB images are shown in Figure 2a, and the collected depth images are shown in Figure 2b. RealSense D455 was used as the acquisition device. The processed depth images are shown in Figure 2c. To better cover all kinds of scenes at night, the dataset of this paper contains pedestrian images taken in different scenes and under different lighting conditions on streets and squares at night. The dataset includes 2226 RGB images and their corresponding depth images.

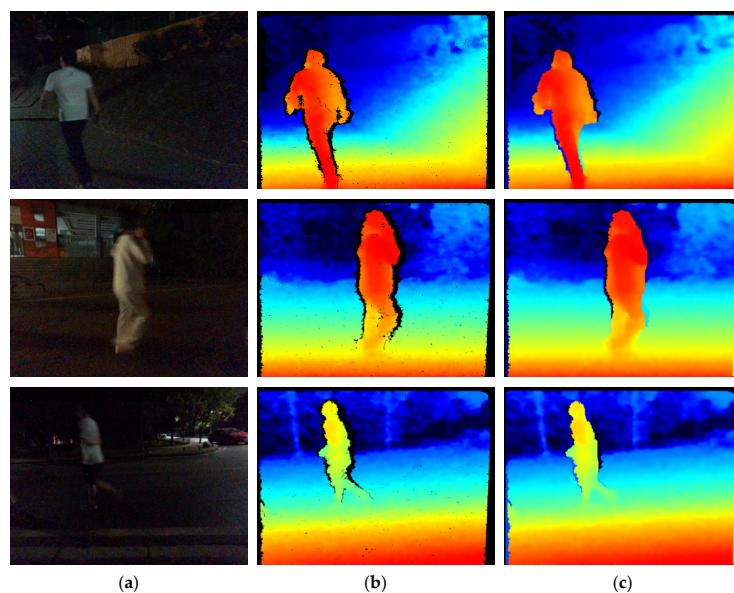


Figure 2. The display of the before and after processing depth images. (a) Original RGB images; (b) Depth images before processing; (c) Depth images after processing.

3.2. Swin-MFA

Swin-MFA is an improved multi-feature fusion network model based on swin-transformer and Unet structure, which retains the basic structure of the encoder and decoder of Unet. The encoder of the Swin-MFA has two input images, namely the depth image and RGB image. During fusion, the two features, respectively, go through feature fusion mechanism, and the final network features are obtained through the attention calculation and the weighted addition. The decoder uses linear layer amplification and rearrangement for up-sampling, and there are residual connections between the encoder and decoder, which can effectively improve the convergence speed of the network. The specific network structure is shown in Figure 3.

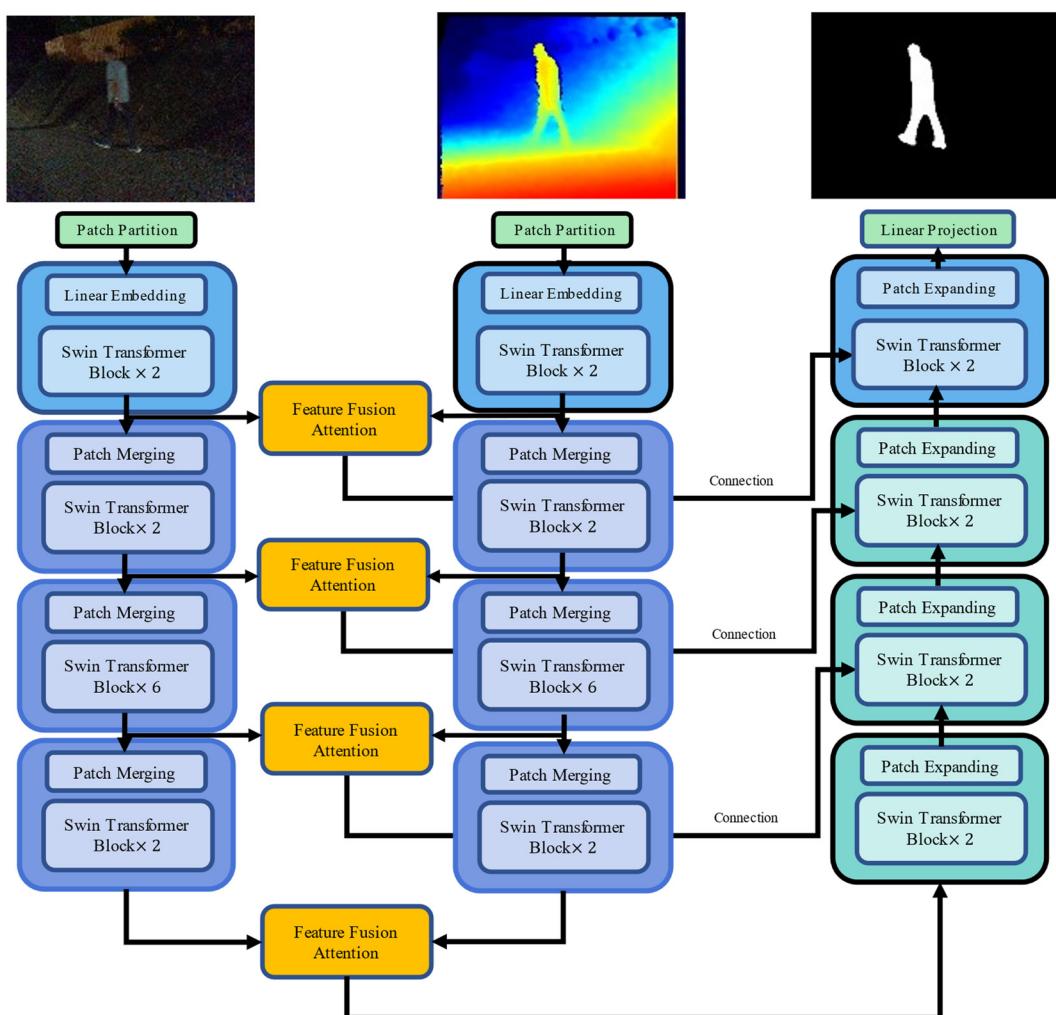


Figure 3. The structure of Swin-MFA network.

3.2.1. Swin-Transformer Base Backbone Network

Due to the condition of low illumination, obvious feature information such as color is missing seriously. It leads to the situation, when extracting features, that the backbone network needs to be insensitive to illumination to ensure the reliability in low-illumination conditions. Therefore, it is meaningful to compare the feature extraction structures of mainstream network structures under different lighting conditions in the same scene. Among them, swin-transformer feature extraction layer performs better than Vision Transformer, ResNet, VGG, MobileNet and the encoder of the Unet structure without additional feature extraction layer in the low illumination conditions. Specific experiments are shown in Figure 4.

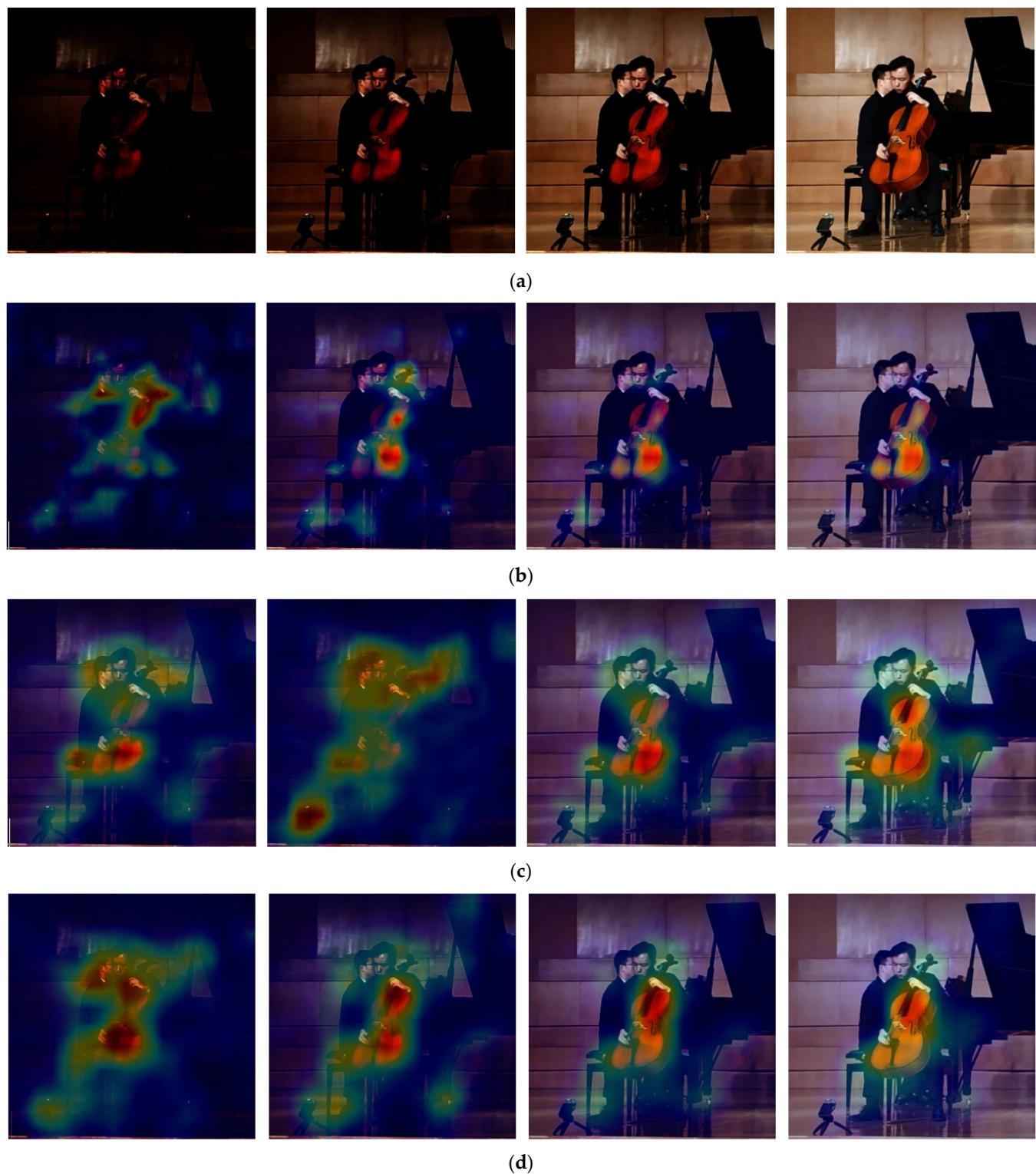


Figure 4. *Cont.*

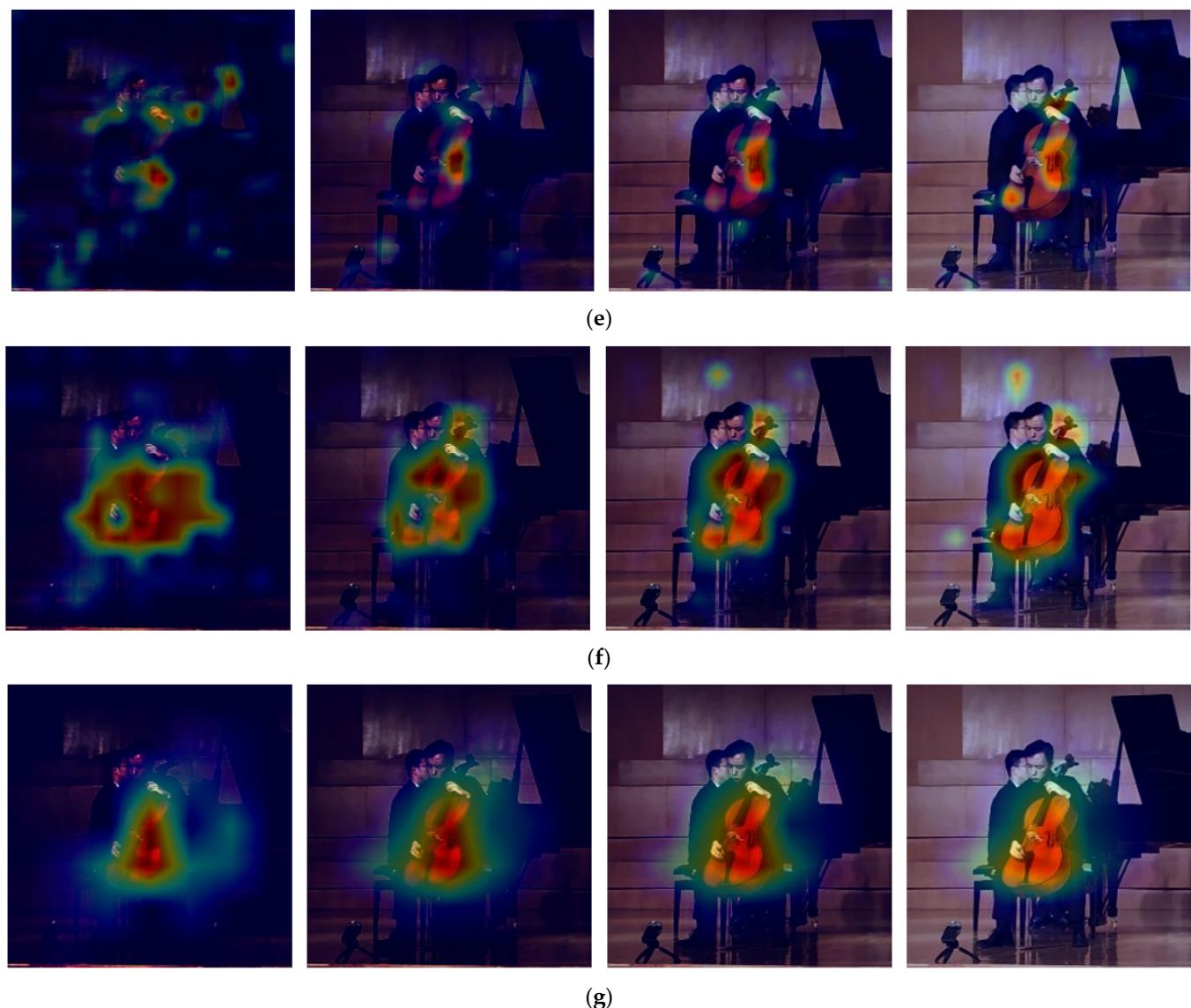


Figure 4. Feature map of different mainstream feature extraction network of cello. (a) Original image of different brightness levels; (b) MobileNet feature map; (c) ResNet-50 feature map; (d) ResNet-101 feature map; (e) VGG feature map; (f) Vision Transformer feature map; (g) swin-transformer feature map.

Using the weights of training on ImageNet, the same image was selected for classification test at different brightness levels, and draw thermal maps by using the grad-CAM [24], which was used to compare the accuracy and concentration of network feature maps at different brightness levels. A score calculation method is defined to measure the feature matching degree of the thermal map on the original image.

$$L = \frac{\sum_{i=1}^M \sum_{j=1}^N G_{ij} Y_{ij}}{\sum_{i=1}^M \sum_{j=1}^N G_{ij}} \quad (5)$$

where, G_{ij} is the two-dimensional output array of grad-CAM. Y_{ij} is the feature matching area of the original image, and, more specifically, it is -1 when it is background and 1 when it is foreground. The score L can be transformed through linear mapping to obtain the result shown in Figure 5.

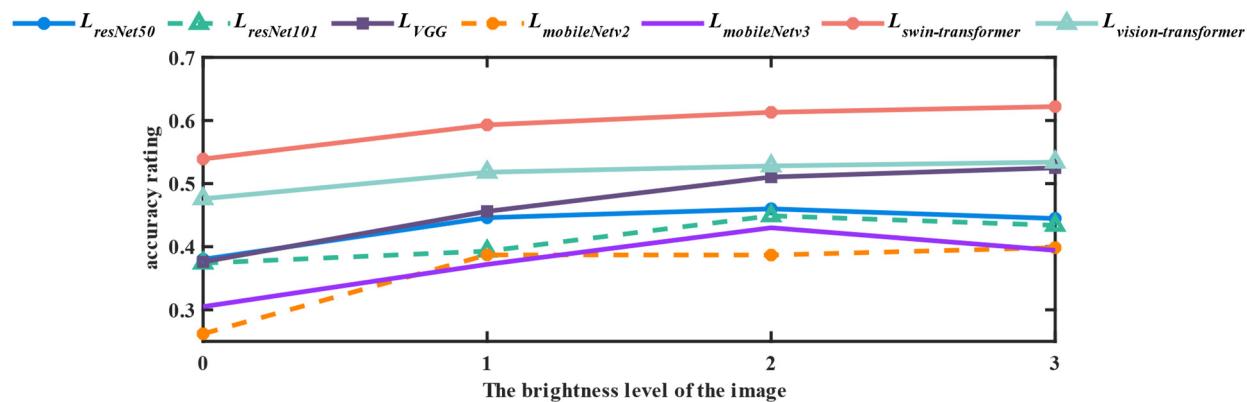


Figure 5. Accuracy diagram of backbone network under different brightness addition.

3.2.2. Self-Attention Mechanism

Swin-transformer's structure contains the form of two multi-headed attention mechanisms, windows multi-head self-attention (W-MSA) and shifted windows multi-head self-attention (SW-MSA). In multi-modal tasks, we also hope to replace the traditional convolutional neural network with the total self-attention. Using the W-MSA module, the network only performs self-attention calculation in windows, and no information is transmitted between windows. Combining with SW-MSA module, the windows slide up to realize information communication between windows, which improves the accuracy and mIoU performance of the network. The specific structure and an illustration of the shifted-window approach are shown in the Figures 6 and 7, respectively.

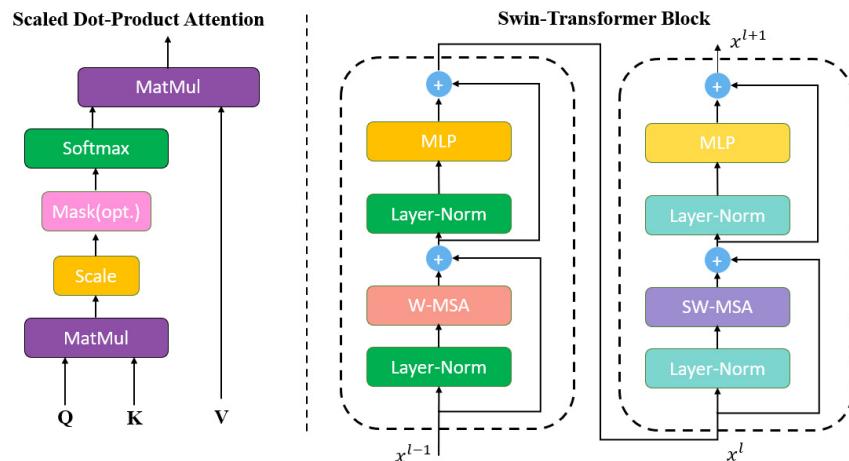


Figure 6. The structure of self-attention mechanism and block in swin-MFA.

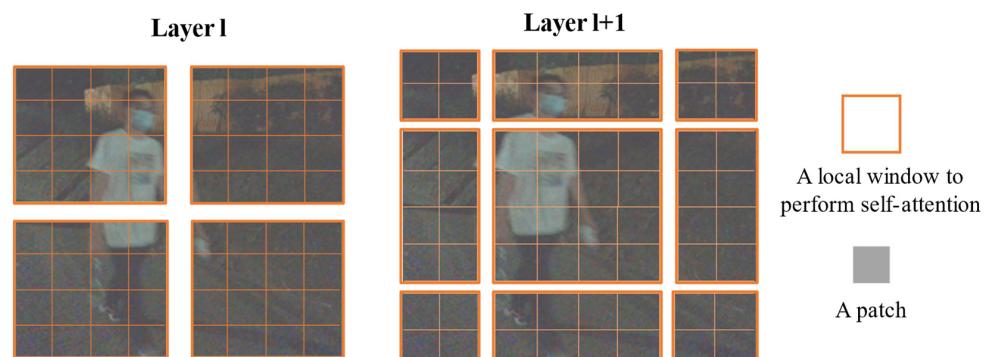


Figure 7. An illustration of the shifted-window approach.

Moreover, in each swin-transformer block, the sliding-window partition mechanism and calculation can be expressed as the equations in (6).

$$\begin{aligned}\hat{x}^l &= W - MSA(LN(x^{l-1})) + x^{l-1} \\ x^l &= MLP(LN(\hat{x}^l)) + \hat{x}^l \\ \hat{x}^{l+1} &= SW - MSA(LN(x^l)) + x^l \\ x^{l+1} &= MLP(LN(\hat{x}^{l+1})) + \hat{x}^{l+1}\end{aligned}\quad (6)$$

where W -MSA and SW -MSA are the formula expression of W-MSA and SW-MSA in Figure 5, respectively. The LN represents the LayerNorm operation. More precisely, the self-attention mechanism can be a query with a series of key-value pairs mapped to a specific output. It can be expressed by Equation (7).

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

where Q K V represents three independent matrices, which are the results of different linear transformations of the original sequence X , and all of these can be used as representatives of X . d_k is the dimension of feature.

3.2.3. Feature-Fusion Attention Mechanism

For the features of the RGB image and depth image generated in the same scene, there are different processing methods on the feature fusion layer. Inspired by ResNet, an additive operation can be used for the feature fusion. Moreover, concatenate operation is used in DenseNet. For confirming the effect of the feature fusion of the RGB and depth images in low illumination conditions, the experiments on addition operation and concatenate operation are conducted, respectively, which can be identified by Figure 8a,b.

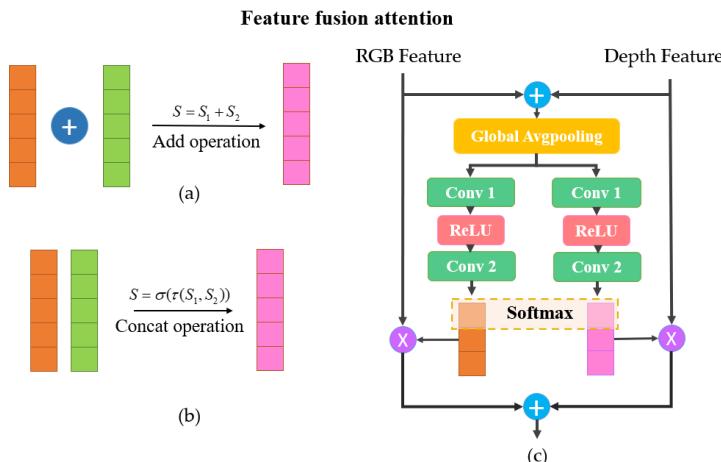


Figure 8. Schematic diagram of image feature fusion operation. (a) The schematic diagram of addition operation; (b) The schematic diagram of concat operation; (c) The schematic diagram of feature-fusion attention mechanism.

The addition operation increases the amount of information that describes the image, but the feature dimension of the image does not change. With more feature sources contained, the increased amount of information in each dimension is obviously beneficial to image segmentation. However, the concatenate operation is the combination of the channels, in other words, the feature dimension of the image has changed. As the dimension of the concatenate operation is increased, the amount of information under each feature has no essential transformation compared with the addition operation. From another perspective, the addition operation is actually a convolution kernel with the corresponding channels sharing the same weight after concatenate.

Inspired by SKnet [25], we designed a fusion attention mechanism for depth features and RGB features. Through reshaping and the global average pooling of the two features, the convolution kernel activation function operation is used to generate the weight matrix of the two channels, respectively. Moreover, the SoftMax processing is carried out in the horizontal dimension of the features by two independent multi-layer perceptrons and, finally, multiplied with the original input; then, the fusion attention of the two features is realized. More specifically, the structure is shown in Figure 8c, and it can be expressed by the Equations (8)–(11).

$$X_c = W_{conv2} \cdot \delta(W_{conv1} \cdot F_{gp}(x_c)) = W_{conv2} \cdot \delta(W_{conv1} \cdot \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W x_c(i, j)) \quad (8)$$

$$Y_c = W_{conv'2} \cdot \delta(W_{conv'1} \cdot F_{gp}(y_c)) = W_{conv'2} \cdot \delta(W_{conv'1} \cdot \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W y_c(i, j)) \quad (9)$$

$$a_c = \frac{e^{X_c}}{e^{X_c} + e^{Y_c}}, b_c = \frac{e^{Y_c}}{e^{X_c} + e^{Y_c}} \quad (10)$$

$$z_c = a_c \cdot x_c + b_c \cdot y_c, c \in [1, C] \quad (11)$$

where x and y are derived from RGB and depth features with the dimensions $H \times W \times C$, respectively. x_c and y_c are the c -th subfeature of x and y . δ is the ReLU activation function, and $z = [z_1, z_2 \dots z_C]$ is the fused attention matrix of the final output. W_{conv} and F_{gp} stand for convolution and global pooling operation.

3.3. Loss Function

The loss function of the network adopts the cross-entropy loss function, which can be expressed by Equation (12). The mask of the loss function calculation is set to ensure the accuracy of itself. At the same time, Adam is used as the optimizer to train on the LLHS dataset proposed in this paper.

$$Loss(p, q) = - \sum_{i=1}^C p_i \log(q_i) \quad (12)$$

where C represents the number of categories, p is the ground truth and q is the predicted result.

4. Results

In this section, based on the results from previous experiments in Section 3.2.1, which prove that the swin-transformer backbone maintains relatively stable feature extraction performance in low-illumination scenes, we performed experiments on the LLHS dataset. In Section 4.1, we compare Swin-MFA with various feature-fusion methods, and the experiment proves that the feature-fusion attention block performs better than other traditional methods. In Section 4.2, ablation experiments were performed in the residual connections between the encoder and decoder network. In Section 4.3, we compare our methods with classic image segmentation methods, such as Lraspp Deeplabv3, HRNet, Trans-Unet, and Swin-Unet, as well as with ACNet, RFNet, 3DGNN, ESANet, FuseNet, CEN, etc. RGB-D multi-modal image segmentation methods are also compared. It shows that our network and multi-modal fusion attention mechanism are effective and reliable. In addition, global acc and mIoU, which are commonly used in image segmentation, are also used to evaluate the results. In addition, more specifically, they can be written by Equations (13) and (14).

$$global\ acc = \frac{\sum_i n_{ii}}{\sum_i N_i} \quad (13)$$

$$mean\ IoU = \frac{1}{n_{class}} \sum_i \frac{n_{ii}}{N_i + \sum_j n_{ji} - n_{ii}} \quad (14)$$

where n_{ij} is the number of pixels with an i -th category that is predicted to be the j -th categories. N_i is the number of total pixels of the i -th category.

4.1. Network Fusion Mechanism Experiments

For the fusion mechanism mentioned in Section 3.2.3, we carried out experiments including addition, concatenate with linear cascade and our feature-fusion attention mechanism. The experimental results are shown in Table 1.

Table 1. The results in the experiment of fusion mechanism.

Fusion Method	Global acc	mIoU
Add	93.4	80.7
Concat	92.8	80.8
Ours	93.6	81.0

It is generally believed that the concatenate operation can cover the addition operations in the effect in improving the segmentation accuracy of the model. However, it is difficult to train the network due to the deep level of the network and the difficulty of convergence. In addition, we notice that the training methods and pre-training weights have an impact on the effect of the feature-fusion methods in the deep network.

4.2. Network Connections between Encoder and Decoder Experiments

Considering the connection forms of encoder and decoder and referring to the way of encoder and decoder of Unet, we verified the function of connections between encoder and decoder after the feature-fusion attention mechanism, which are tested in the situations of no connection, single connection and multiple connections, respectively. The specific results are shown in Table 2.

Table 2. The results in the experiment of the number of connections.

Number of Connections	Global acc	mIoU
0	89.2	73.7
1	91.8	76.6
2	92.3	78.8
3	93.4	81.0

4.3. Network Comparative Experiments

We compared our methods with Lraspp, Deeplabv3, TransUnet, SwinUnet, ACNet, RFNet, 3DGNN ESANet, FuseNet, LDFNet, etc. The specific results are shown in Table 3 and Figure 9. The experimental results show that our method is effective and accurate.

Table 3. The mIoU and global accuracy results of the comparison experiment.

Method	Global acc	mIoU
Lraspp [26]	73.5	63.4
Deeplabv3 [27]	84.1	54.3
Unet [11]	81.0	71.7
HRNet [28]	46.7	59.1
TransUnet [29]	88.0	75.5
SwinUnet [30]	87.5	69.8
ACNet [23]	92.7	75.7
RFNet [31]	82.4	72.3
3DGNN [22]	92.2	77.6
ESANet [32]	92.2	80.4
FuseNet [33]	84.3	75.3
RedNet [34]	88.5	75.0
LDFNet [21]	89.9	78.3
Swin-MFA (Ours)	93.4	81.0

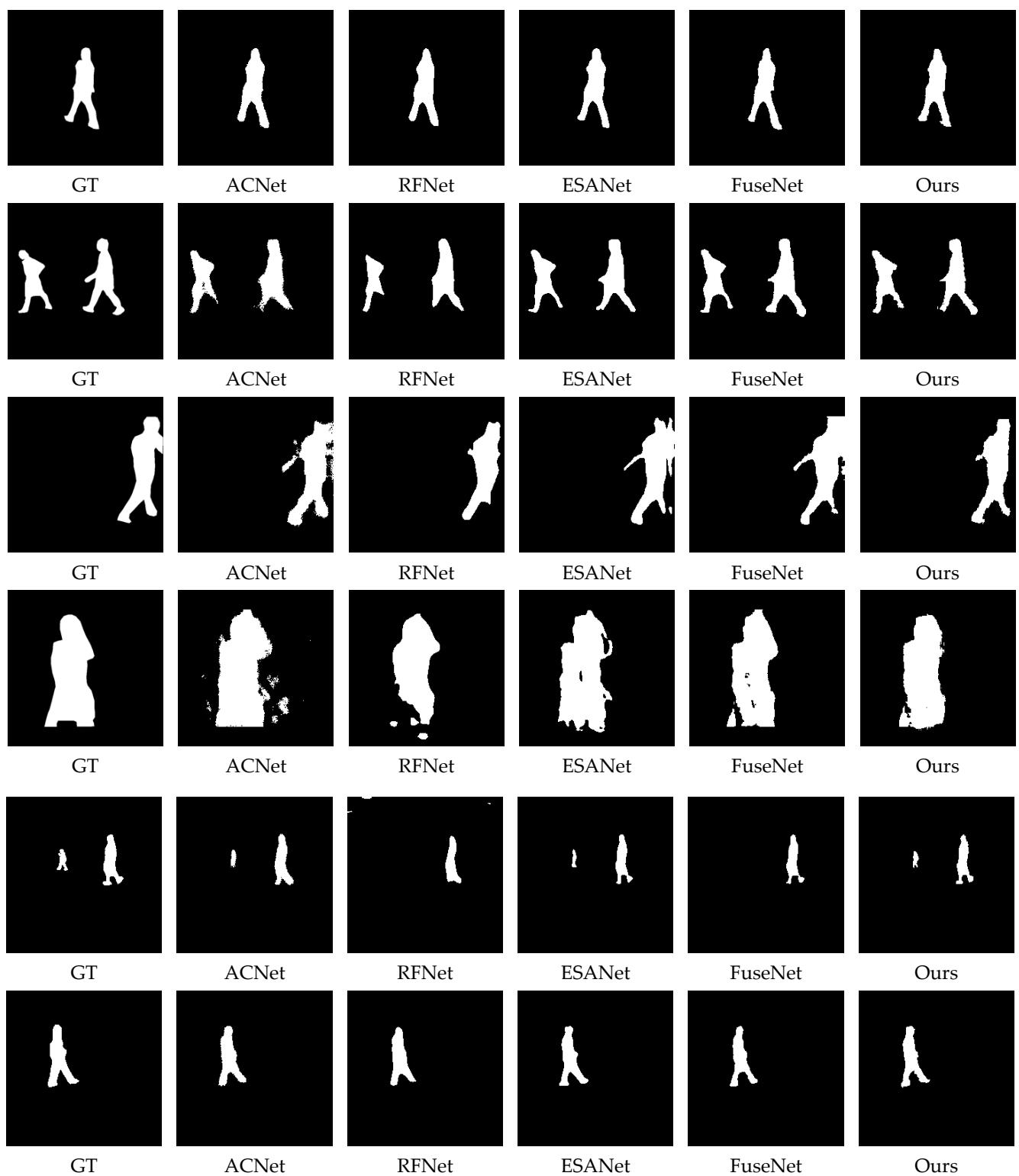


Figure 9. Experimental results compared with other methods.

4.4. Experiment of the Combining Datasets of Different Light Intensities

For the low-illuminance monitoring at night, there are occasional lights such as car lights, so it is necessary to conduct data analysis on different brightness conditions. We added five levels of high and low brightness mixing datasets to the LLHS dataset, with high brightness accounting for 10%, 15%, 20%, 25% and 30%, respectively. The specific results are shown in Table 4.

Table 4. The results of experiment in datasets of different light levels.

Method	10%		15%		20%		25%		30%	
	mIoU	acc								
ACNet	75.6	92.6	75.0	92.4	75.2	92.0	75.1	92.5	75.1	92.1
RFNet	72.3	82.7	71.5	82.3	71.5	81.8	71.3	82.4	72.1	82.5
3DGNN	77.7	92.5	77.4	92.5	77.1	92.0	77.3	92.1	76.9	91.1
ESANet	80.3	92.2	80.0	92.1	79.8	91.7	80.6	92.2	79.9	91.8
FuseNet	75.1	84.2	74.0	84.4	74.2	84.0	75.5	86.4	74.6	84.7
RedNet	74.8	88.3	71.8	88.4	72.9	87.6	70.6	88.5	73.4	87.4
Swin-MFA	80.8	92.9	80.6	92.9	80.1	92.4	80.5	92.6	80.2	92.9

5. Discussion

We demonstrate the robustness of a swin-transformer network in low-illumination conditions through comparative experiments and introduce a total self-attention mechanism to replace the traditional convolutional neural network, to improve the ability of the model's attention to depth images and RGB images. Moreover, a fusion attention mechanism is proposed, to make the overall network have better performance. At present, we are implementing semi-supervised learning on the network, which has made preliminary progress. In the future, we will continue to expand the performance of the network and the active learning ability on datasets without a label.

6. Conclusions

In this paper, an end-to-end multi-modal image segmentation transformer network is proposed. Through the multi-modal fusion attention of the depth images and RGB images, the human-segmentation problem in the conditions of low illumination is solved, which can be well-applied in the monitoring and security fields. Depth image and RGB image were used as complementary inputs, and the neural network structure of the multi-modal encoder and decoder was used to realize the segmentation task in complex low-illumination conditions, which improved the robustness and learning performance of the network. In addition, we first propose a low-illuminance human-segmentation dataset, which fills the gap of the multi-modal low-illuminance dataset. Experimental results show that the proposed method is far superior to the advanced single-mode segmentation method as well as the depth and RGB multi-modal network method, with better performance in low-illumination conditions. In the future, we will also realize semi-supervised and unsupervised active-learning strategies by the network, so that the network can still have excellent performance without accurate annotation.

Author Contributions: X.Y. and H.Z. designed the study; X.Y. and Y.W. contributed to the methodology; X.Y., H.Z., Y.W. and J.W. performed the experiments; H.Z. contributed to the data preprocessing; X.Y. and H.Z. analyzed the data; X.Y. and S.G. wrote the paper and created the diagrams; C.F. and J.W. contributed to the advising and proofreading. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the innovation and entrepreneurship research project for the colleges, China: 202110486061.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the privacy issues that the data contained a large number of portraits and pictures of the participants, who did not agree to be posted directly on the Internet. But some applications for research with reasonable requests are still allowed to use.

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

References

1. Liu, X.; Deng, Z.; Yang, Y. Recent progress in semantic image segmentation. *Artif. Intell. Rev.* **2019**, *52*, 1089–1106. [[CrossRef](#)]
2. Wang, W.; Zhou, T.; Yu, F.; Dai, J.; Konukoglu, E.; Van Gool, L. Exploring cross-image pixel contrast for semantic segmentation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 7283–7293.
3. Nobuyuki, O. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern.* **1979**, *9*, 62–66.
4. Yang, L.; Wu, X.Y.; Zhao, D.W.; Li, H.; Zhai, J. An improved Prewitt algorithm for edge detection based on noised image. In Proceedings of the 2011 4th International Congress on Image and Signal Processing, Shanghai, China, 15–17 October 2011; pp. 1197–1200.
5. Coates, A.; Ng, A.Y. Learning feature representations with K-means. *Lect. Notes Comput. Sci.* **2012**, *7700*, 561–580.
6. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [[CrossRef](#)] [[PubMed](#)]
7. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [[CrossRef](#)] [[PubMed](#)]
8. Wu, X.Y.; Wu, Z.Y.; Guo, H.; Ju, L.L.; Wang, S. DANNet: A One-Stage Domain Adaptation Network for Unsupervised Nighttime Semantic Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 15764–15773.
9. Sun, Y.X.; Zuo, W.X.; Liu, M. RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes. *IEEE Robot. Autom. Lett.* **2019**, *4*, 2576–2583. [[CrossRef](#)]
10. Chen, X.; Lin, K.Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; Zeng, G. Bi-directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 561–577.
11. Ronneberger, O.; Fisher, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
12. Zhao, H.S.; Qi, X.J.; Shen, X.Y.; Shi, J.; Jia, J. Icnet for real- time semantic segmentation on high-resolution images. In Proceedings of the 2018 European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 418–434.
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
14. Lin, T.Y.; Wang, Y.X.; Liu, X.Y.; Qiu, X. A Survey of Transformers. *arXiv* **2021**, arXiv:2106.04554.
15. 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* **2020**, arXiv:2010.11929.
16. Liu, Z.; Lin, Y.T.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
17. He, X.; Chen, Y.S.; Lin, Z.H. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 498. [[CrossRef](#)]
18. Lu, Y.T.; Fu, J.; Li, X.; Zhou, W.; Liu, S.; Zhang, X.; Jia, C.; Liu, Y.; Chen, Z. RTN: Reinforced Transformer Network for Coronary CT Angiography Vessel-level Image Quality Assessment. *arXiv* **2022**, arXiv:2207.06177.
19. Zheng, S.X.; Lu, J.C.; Zhao, H.S.; 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; pp. 6877–6886.
20. Schneider, L.; Jasch, M.; Frohlich, B.; Weber, T.; Franke, U.; Pollefey, M.; Ratsch, M. Multimodal neural networks: RGB-D for semantic segmentation and object detection. *Lect. Notes Comput. Sci.* **2017**, *10269*, 98–109.
21. Hung, S.W.; Lo, S.Y.; Hang, H.M. Incorporating Luminance, Depth and Color Information by a Fusion-based Network for Semantic Segmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, China, 22–25 September 2019; pp. 2374–2378.
22. Qi, X.J.; Liao, R.J.; Jia, J.Y.; Fidler, S.; Urtasun, R. 3D Graph Neural Networks for RGBD Semantic Segmentation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5209–5218.
23. Hu, X.; Yang, K.; Fei, L.; Wang, K. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1440–1444.
24. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [[CrossRef](#)]
25. Li, X.; Wang, W.H.; Hu, X.L.; Yang, J. Selective Kernel Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
26. Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324.

27. 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](#)]
28. Sun, K.; Zhao, Y.; Jiang, B.R.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
29. Chen, J.N.; Lu, Y.Y.; Yu, Q.H.; Luo, X.D.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y.Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
30. Cao, H.; Wang, Y.Y.; Chen, J.; Jiang, D.S.; Zhang, X.P.; Tian, Q.; Wang, M.N. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv* **2021**, arXiv:2105.05537.
31. Sun, L.; Yang, K.; Hu, X.; Hu, W.; Wang, K. Real-Time Fusion Network for RGB-D Semantic Segmentation Incorporating Unexpected Obstacle Detection for Road-Driving Images. *IEEE Robot. Autom. Lett.* **2020**, *5*, 5558–5565. [[CrossRef](#)]
32. Seichter, D.; Kohler, M.; Lewandowski, B.; Wengfeld, T.; Gross, H.-M. Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13525–13531.
33. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. In Proceedings of the 13th Asian Conference on Computer Vision (ACCV), Taipei, Taiwan, 20–24 November 2016; pp. 213–228.
34. Jiang, J.D.; Zheng, L.N.; Luo, F.; Zhang, Z. RedNet: Residual Encoder-Decoder Network for indoor RGB-D Semantic Segmentation. *arXiv* **2018**, arXiv:1806.01054.