Pattern Recognition 144 (2023) 109862

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Cross-Modal Transformer for RGB-D semantic segmentation of production workshop objects

Qingjun Ru, Guangzhu Chen *, Tingyu Zuo, Xiaojuan Liao

College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, China

ARTICLE INFO

Keywords:
Cross-Modal
Production workshop object
RGB-D
Semantic segmentation
Transformer

ABSTRACT

Scene understanding in a production workshop is an important technology to improve its intelligence level, semantic segmentation of production workshop objects is an effective method for realizing scene understanding. Since the varieties of information of production workshop, making full use of the complementary information of RGB image and depth image can effectively improve the semantic segmentation accuracy of production workshop objects. Aiming at solving the multi-scale and real-time problems of segmenting the production workshop objects, this paper proposes Cross-Modal Transformer (CMFormer), a Transformer-based cross-modal semantic segmentation model. Its key feature correction and feature fusion parts are composed of the Multi-Scale Channel Attention Correction(MS-CAC) module and the Global Feature Aggregation(GFA) module. By improving Multi-Head Self-Attention(MHSA) in Transformer, we design Cross-Modal Multi-Head Self-Attention(CM-MHSA) to build long-range interaction between RGB image and depth image, and further design the MS-CAC module and the GFA module on the basis of the CM-MHSA module, to achieve cross-modal information interaction in the channel and spatial dimensions. Among them, the MS-CAC module enriches the multi-scale features of each channel and achieve more accurate channel attention correction between the two modals; the GFA module interacts with RGB feature and depth feature in the spatial dimension and fuses global and local features at the same time. In the experiments on the NYU Depth v2 dataset, the CMFormer reached 68.00% MPA(Mean Pixel Accuracy) and 55.75% mIoU(Mean Intersection over Union), achieves the state-of-the-art results. While in the experiments on the Scene Objects for Production workshop dataset(SOP), the CMFormer achieves 96.74% MPA, 92.98% mIoU and 43 FPS(Frames Per Second), which has high precision and good real-time performance. Code is available at: https://github.com/FutureIAI/CMFormer

1. Introduction

With the development of industrial internet technology, the traditional manufacturing industry is developing towards intelligent manufacturing. As an important part of the manufacturing industry, accurate understanding of the workshop scene is an important part of realizing intelligent manufacturing. The scene understanding technology can make the equipment accurately perceive the workshop operation situation, speed up the workshop operation efficiency, and improve the production and manufacturing capacity.

Image semantic segmentation, as an effective scene understanding technique, aims to assign a class label to each pixel in an image and predict the location and shape of an object. Semantic level perceptual recognition of workshop scene objects is the foundation for achieving workshop intelligence, such as workshop intelligent security and mobile robot intelligent navigation tasks. The above tasks require semantic level perception and recognition of workshop scene objects, that is, identifying the type, shape, and pose of the objects, and then making inference decisions based on the recognition results. However, most semantic segmentation studies are based on single-modal of RGB image [1,2], which are difficult to adapt to environments with high complexity and uneven illumination, and cannot achieve accurate understanding of the production workshop scene. With the wide application of depth camera, depth information of an image has used to assist semantic segmentation [3]. Introducing the depth information into the RGB image as a supplement is more conducive to distinguish the confused areas in an image. Therefore, it is of great significance to study RGB-D semantic segmentation methods in production workshop scene.

However, how to find the fusion modals of RGB feature and depth feature is always a challenging problem. FCN [4] simply took the depth

* Corresponding author.
E-mail address: cgzhu@126.com (G. Chen).

https://doi.org/10.1016/j.patcog.2023.109862
Received 12 July 2022; Received in revised form 14 July 2023; Accepted 31 July 2023
Available online 1 August 2023

image as the fourth channel of RGB image, and spliced it with RGB image as the input of Convolutional Neural Network(CNN). While other methods [5,6] used two-stream CNN as an encoder to extract features from RGB image and depth image respectively, the two modal features of multiple stages were fused, and finally the fused features were decoded to obtain the final segmentation result. With the emergence of the attention mechanism [7,8], some methods [9–11] utilized the attention mechanism to balance the distribution of two modal features, made the network pay more attention to the effective regions of the image.

Recently, Transformer structure [12], which models the global relations through self-attention mechanism, has achieved great success in the field of Natural Language Processing (NLP). Inspired by this significant achievement, many researchers have applied Transformer structure to computer vision field and proved the effectiveness of this structure in various computer vision tasks. Some methods [13–15] have obtained better results than CNN in the field of RGB-D semantic segmentation using the cross-modal characteristics of Transformer.

Nevertheless, the aforementioned based-CNN RGB-D semantic segmentation methods more focused on the local features of an image without considering the long-range dependency information. At the same time, the above RGB-D semantic segmentation methods based on attention mechanism were implemented by using the Channel Attention Mechanism(CAM) and the Spatial Attention Mechanism(SAM) proposed in SENet [7] and CBAM [8]. Among them, the CAM uses the global pooling method to obtain the channel attention vector, which will cause the image to lose too many features, especially multi-scale features; the SAM calculates the spatial attention vector through convolution operation, but the global receptive field cannot be obtained by stacking several convolutional layers, so the global features are ignored. In the Transformer-based RGB-D semantic segmentation methods, CMX [14] and UCTNet [16] also use the original the CAM in feature correction and feature fusion without improving it. TransD-Fusion [13] only performs feature fusion in the input part and the down-sampling part of the final stage, without using multi-level fusion method. Therefore, the above calculation methods of attention mechanism have certain limitations.

This paper aims to propose a new feature correction module and feature fusion module based on pure Transformer and apply it to the production workshop scene to achieve certain application value in the field of intelligent manufacturing. Since the production workshop scene is a complex scene, the implementation of RGB-D semantic segmentation mainly faces the following challenges:

(1) In the production workshop, there are large scale differences between different objects, so how to make full use of multi-scale features to segment the objects in the production workshop scenes is a key problem.
(2) In the production workshop, various equipment needs to perceive the surrounding environment in real time, so how RGB-D semantic segmentation does meet real-time performance under high accuracy is another key problem.

In order to overcome the above challenges and the limitations of the existing attention mechanism, we propose an RGB-D semantic segmentation method based on Transformer, which is called Cross-Modal Transformer(CMFormer). The CMFormer follows the standard encoder-decoder structure, including a two-stream feature extraction encoder and a decoder. We designed the Cross-Modal Multi-Head Self-Attention(CM-MHSA) by improving the Multi-Head Self-Attention (MHSA) in the original Transformer [12], achieving long-range information interaction between RGB image and depth image. On the basis of CM-MHSA, we further designed the Multi-Scale Channel Attention Correction(MS-CAC) module and the Global Feature Aggregation(GFA) module on the basis of the CM-MHSA module, to achieve cross-modal information interaction in the channel and space dimensions. Among them, the MS-CAC module is a new channel attention module which is different from the previous RGB-D semantic segmentation methods, which combines SPPNet [17] and Transformer [12] to enrich the multi-scale features of each channel and achieve more accurate channel attention correction between the two modals; the GFA module interacts with RGB feature and depth feature in the spatial dimension and fuses global and local features at the same time, and then inputs the fused features into the decoder. Compared with the existing RGB-D semantic segmentation methods, the CMFormer has stronger global modeling ability, and better abilities of information interaction and fusion in channel and spatial dimensions.

In summary, the main contributions of this paper can be summarized as follows:

(1) A new RGB-D semantic segmentation method based on Transformer(CMFormer) is proposed. Its key feature correction and feature fusion parts are composed of the MS-CAC module and the GFA module. The MS-CAC module preserves multi-scale information while realizing channel feature correction, while the GFA module realizes spatial feature fusion, and fully considers global and local features, which make the CMFormer achieve better cross-modal information interaction by capturing long-range contextual dependencies.
(2) The effectiveness of the CMFormer is evaluated with extensive experiments on the SOP dataset [18] and NYU Depth v2 dataset [19]. The results show that the CMFormer achieves the state-of-the-art results on both datasets. In the experiment of SOP dataset [18], CMFormer can achieve 96.74% MPA and 92.98% mIoU. At the same time, the real-time performance of CMFormer can reach 43 FPS(Frames Per Second), which has good real-time performance while meeting the requirements of high precision.

2. Related work

2.1. Vision transformer

2.1.1. Transformer backbones

Vision Transformer(ViT) [20] applied a pure Transformer structure into image classification task for the first time. When the data is large enough(i.e. on ImageNet-21k [21], JFT-300M [22]), the classification performance of ViT is close to or even better than that of CNN. Specifically, ViT splits the image into 16×16 patches, and stacks multiple standard Transformer layers to make the image classification. Subsequently, DeiT [23] proposed an efficient training method and a teacher-student strategy specific to Transformer to make ViT get rid of the limitation of dataset. Beside, T2T-ViT [24], TNT [25], and CrossViT [26] made tailored changes to ViT [20] for different problem angles to further improve the performance of image classification.

In order to improve the performance of Transformer on dense prediction tasks, PVT [27] introduced a pyramid structure in Transformer for the first time, and extended the PVT [27] to other vision tasks. CvT [28] and CeiT [29] and introduced CNN into Transformer, which not only obtained the desirable properties of CNN, but also maintained the advantages of Transformer. CPVT [30] replaced previous fixed length positional encodings with conditional positional encoding(CPE), enabling Transformers to process images of different sizes without interpolation. Swin Transformer [31] used a hierarchical construction method to make the model have a pyramid structure. At the same time, the shifted window operation was introduced to limit self-attention in windows, which greatly reduced the amount of calculation.

2.1.2. Transformer in Semantic Segmentation

SETR [32] replaced the CNN backbone with ViT [20], transformed semantic segmentation into a sequence-to-sequence prediction task, resulting in better feature representation. SegFormer [33] designed Efficient Self-Attention and Mix-FFN to improve Transformer for semantic segmentation tasks and redesigned a lightweight decoder with

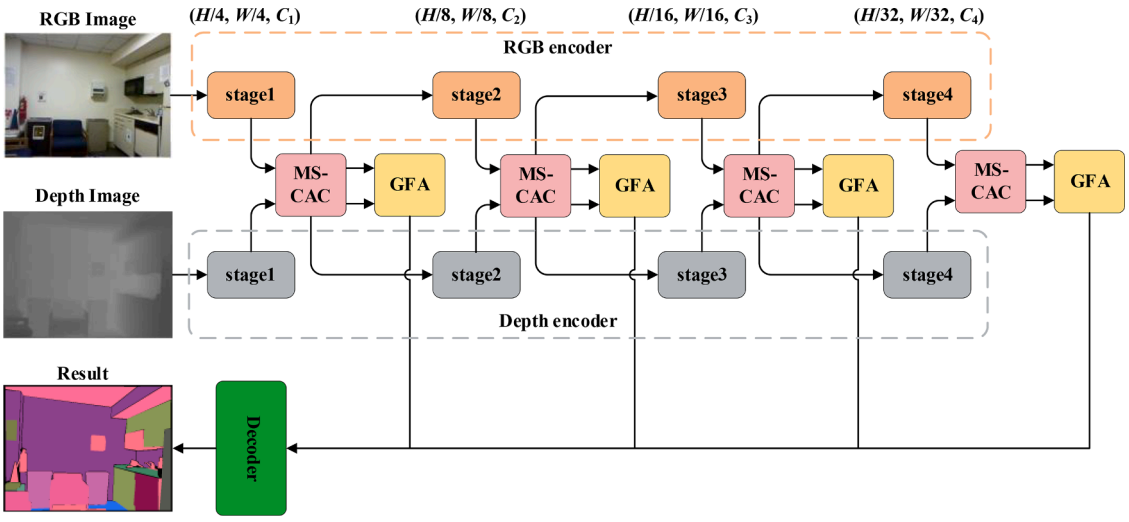


Fig. 1. Overall architecture of the CMFormer. The CMFormer mainly consists of three parts: (1) Two-stream feature extraction encoder is used to extract RGB feature and depth feature respectively, where stage1-4 represent each down-sampling stage of different modal encoders. (2) The cross-modal information interaction part consists of the MS-CAC module and the GFA module, which are used to correct and aggregate different modal features. (3) The decoder is used to fuse features from different stages to generate the final segmentation result.

only four MLP layers. EX-ViT [34] propose a novel Explainable Multi-Head Attention and the Attribute-guided Explainer to provides model-inherent explainable attention maps and recognize discriminative attribute features for a target object with image level labels. Segmenter [35] adopted ViT [20] in the encoding stage, split the image into patches, and outputted the embedded sequences after being processed by the encoder. In decoding stage, the class labels were obtained from these embedded sequences with a point-wise linear decoder or a mask Transformer decoder.

2.2. RGB-D semantic segmentation

With the wide application of depth sensors, we can more easily obtain the depth information of a scene. An effective fusion method of RGB feature and depth feature can improve the accuracy of semantic segmentation. Some early studies have manually designed feature description methods to improve the performance of RGB-D semantic segmentation [36,37]. With the wide application of CNN in the field of computer vision, CNN gradually occupied a dominant position in RGB-D semantic segmentation. FCN [4] simply concatenated the RGB channel and the depth channel of the image as the four-channel input of CNN. However, due to less meaningful gradient information transmitted in the training process, the depth information as an input channel does not bring significant performance gains to semantic segmentation models. Gupta et al. [38] proposed HHA(Horizontal disparity, Height above ground, Angle of the surface normal) depth information representation method, which converted the depth image into three different channels, and then RGB image and HHA image were respectively inputted into CNN for feature extraction, and fused in the final stage of the network. FUSENet [39] used a two-stream CNN encoder to extract features from RGB image and depth image respectively, and selectively fused different levels of depth features into the corresponding RGB features to achieve RGB-D semantic segmentation. RDFNet [6] applied the core idea of residual learning to effectively extract RGB features and complementary depth features, and used the way of thermocline connection to learn multi-level fusion features, so as to obtain better semantic segmentation prediction results.

ACNet [9] used the CAM to filter important features in RGB image and depth image, and fuses features between different modals through a three parallel branches structure. Both SA-Gate [10] and CMX [14] use the method of combining the CAM and SAM to achieve feature correction between different modals. The difference is that SA-Gate [10] achieves feature fusion by adding, while CMX uses MHSA to achieve feature fusion. CMX [14] used Efficient Attention [40] to reduce the huge computational cost in the process of calculating self -attention. The TransD-Fusion [13] achieved the final segmentation effect through three stages of self-enhancement, cross-calibration and depth-guided fusion in the final stage of down-sampling. To improve the effect of feature extraction, UCTNet [16] further considered the imaging quality of depth image, and proposed an Uncertainty-Aware Self-Attention mechanism to limit the information flow of depth image.

3. Cross-Modal Transformer

In this section, we first introduce the overall architecture of the CMFormer in Section 3.1, next the CM-MHSA module for cross-modal information interaction is introduced in Section 3.2. The MS-CAC module and the GFA module for cross-modal feature correction and aggregation are introduced in Section 3.3 and Section 3.4, respectively. Finally, a series of the CMFormer models with different sizes, namely The CMFormer-Small, The CMFormer-Medium and The CMFormer-Large, are introduced in Section 3.5.

3.1. Overall architecture

In order to realize the information interaction between RGB feature and depth feature, inspired by ACNet [9], the overall architecture of the CMFormer adopts a standard encoder-decoder structure.

We use two parallel encoders to extract the latent features of the RGB image and depth image, and the decoder decodes the extracted features and outputs the final segmentation result. The overall structure of the model is shown in Fig. 1. The acquisition data of the RGB image and depth image often contain some noises. For the depth image, due to the limited imaging distance of the depth camera, the depth values are inaccurate. For the RGB image, the boundaries of some objects with similar colors and textures usually cannot be distinguished better. These noises have a certain impact on the accuracy of semantic segmentation. During the coding phase, an efficient cross-modal information interaction method can fully utilize the effective information of two modals and identify their strengths from each modal. Therefore, this paper proposes the CM-MHSA mechanism, and designs the MS-CAC module on this basis. The MS-CAC module corrects the features of the two modals in each down-sampling stage, supplements the features of the other modal with the features of one modal, and inputs the output results to the next

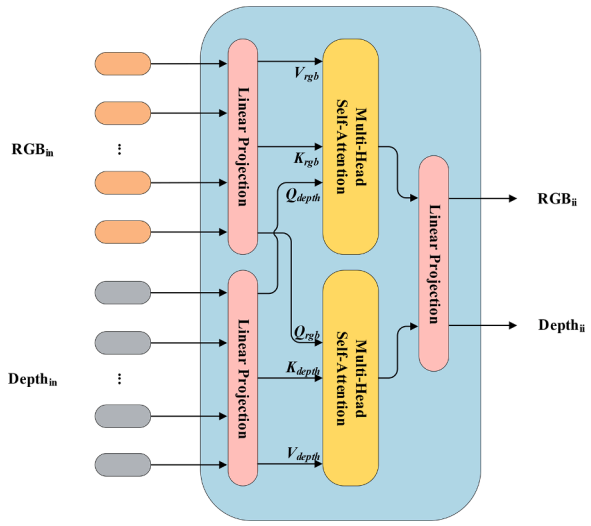


Fig. 2. Cross-Modal Multi-Head Self-Attention (CM-MHSA).

stage. In addition, we also designed the GFA module, which realizes the information exchange in the spatial dimension, and aggregates the feature maps of the two modals into a single feature map through the fusion of global features and local features.

3.2. Cross-Modal Multi-Head Self-Attention module

In order to achieve better cross-modal information exchange, we designed the CM-MHSA module. The CM-MHSA module fully extracts important features of two modals by exchanging Query vectors.

The details of CM-MHSA are shown in Fig. 2. First, the input RGB feature and depth feature dimension are convert from the input of $RGB_{in} \in \mathbb{R}^{H\times W\times C}$, $Depth_{in} \in \mathbb{R}^{H\times W\times C}$to $RGB_{in} \in \mathbb{R}^{N\times C}$, $Depth_{in} \in \mathbb{R}^{N\times C}$ tokens, where N represents the number of tokens, and C represents the dimension of each token. The specific division process will be introduced in detail in Sections 3.3 and 3.4. Then Query Q, Key K, and Value V of the two modals are calculated by the projection matrices W^Q_{RGB}, W^K_{RGB}, W^V_{RGB} and W^Q_{Depth}, W^K_{Depth}, W^V_{Depth}, denoted as Q_{RGB}, K_{RGB}, V_{RGB} and Q_{Depth}, K_{Depth}, V_{Depth}. Different from the original MHSA mechanism, we exchange Q of the two modals, and use Q of one modal and K of the other modal to calculate the self-attention matrix, to achieve cross-modal information interaction, the calculated self-attention matrix is expressed as $Attention_{RGB}$ and $Attention_{Depth}$. Then multiply the obtained self-attention matrix with the V of each mode, and finally obtain the RGB feature RGB_{ii} and depth feature $Depth_{ii}$ after information interaction through a Linear Projection operation, where ii represents information interaction. The above steps can be formulated as follows:

$$Q_{RGB},\ K_{RGB},\ V_{RGB} = RGB_{in}W^Q_{RGB},\ RGB_{in}W^K_{RGB},\ RGB_{in}W^K_{RGB} \tag{1}$$

$$Q_{Depth},\ K_{Depth},\ V_{Depth} = Depth_{in}W^Q_{Depth},\ Depth_{in}W^K_{Depth},\ Depth_{in}W^K_{Depth} \tag{2}$$

$$Attention_{RGB} = Softmax\left(\frac{Q_{Depth}K^T_{RGB}}{\sqrt{d_{head}}}\right), \tag{3}$$

$$Attention_{Depth} = Softmax\left(\frac{Q_{RGB}K^T_{Depth}}{\sqrt{d_{head}}}\right), \tag{4}$$

$$RGB_{ii} = LP(Attention_{RGB}V_{RGB}), \tag{5}$$

$$Depth_{ii} = LP\left(Attention_{Depth}V_{Depth}\right), \tag{6}$$

where d_{head} represents the dimension of each head, LP represents Linear Projection, and T represents matrix transpose operation.

3.3. Multi-Scale Channel Attention Correction module

In an image, the features contained in each channel are different. SENet [7] given each channel a different weight by calculating the channel attention vector to achieve better feature extraction effect. However, the CAM is implemented through the global pooling layer and the fully connected layer, which will lose the multi-scale features in the image. In practical scenes, multi-scale features are the key to distinguishing objects of different scales. In response to the above problem, we propose the MS-CAC module, a new channel attention calculation method. The module is divided into two parts: Image to Tokens part and Channel Attention Correction part. The detailed structure of the MS-CAC module is shown in Fig. 3.

3.3.1. Image to Tokens(I2T)

The purpose of the I2T part is to convert the feature of each channel of the image into a one-dimensional token, so that the shape of the image is converted from $H \times W \times C$ to $C \times C_1$, to satisfy the input form of the Transformer, where C represents the number of channels of the image and C_1 represents the dimension size of each token.

First, the input features $RGB_{in} \in \mathbb{R}^{H\times W\times C}$ and $Depth_{in} \in \mathbb{R}^{H\times W\times C}$ of the two modals are first down-sampled through a multi-scale pooling operation. We use four different pooling factors to capture the multi-scale features of the image, and the four pooling factor sizes are set to $M_1 \times N_1$, $M_2 \times N_2$, $M_3 \times N_3$, $M_4 \times N_4$. At the same time, the average pooling and max pooling operations are both used to preserve richer information. The pooling result of size $\mathbb{R}^{M\times N\times C}$ is flattened to $\mathbb{R}^{C\times MN}$ and merged along the second dimension to get the pooled features $RGB_{mp} \in \mathbb{R}^{C\times L}$, $RGB_{ap} \in \mathbb{R}^{C\times L}$, $Depth_{mp} \in \mathbb{R}^{C\times L}$, $Depth_{ap} \in \mathbb{R}^{C\times L}$, where

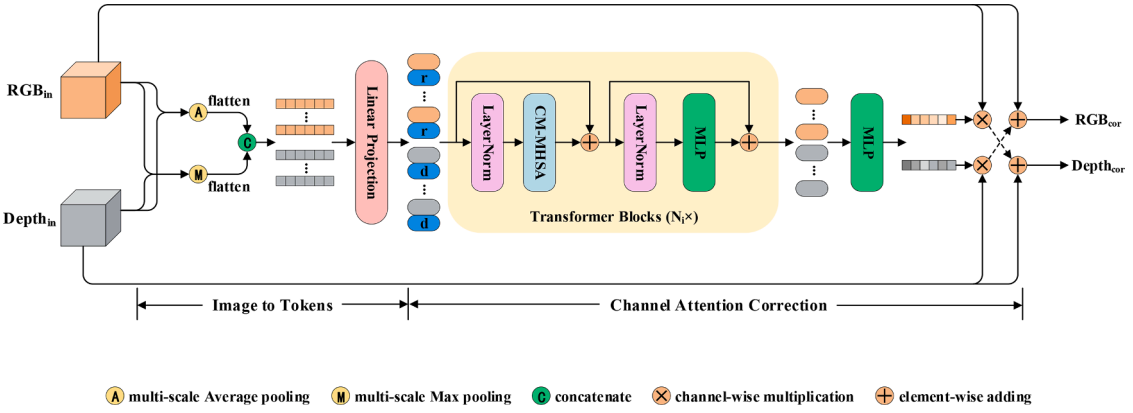


Fig. 3. Multi-Scale Channel Attention Correction.

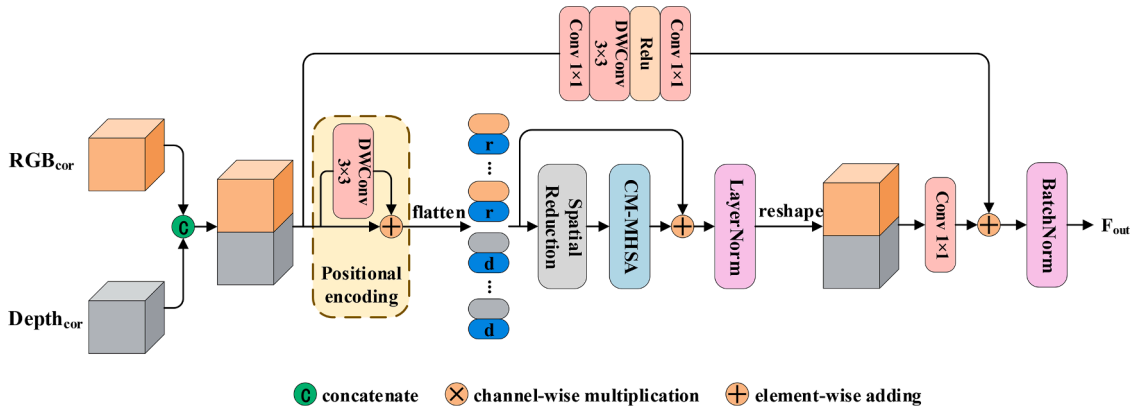


Fig. 4. Global Feature Aggregation.

$L = M_1N_1 + M_2N_2 + M_3N_3 + M_4N_4$, subscripts *mp* and *ap* represent max pooling and average pooling, respectively.

Then, we merge the multi-scale features of the same modal along the second dimension to obtain the multi-scale features of the two modals: $RGB_{msf} \in \mathbb{R}^{C\times 2L}$ and $Depth_{msf} \in \mathbb{R}^{C\times 2L}$, where subscript *msf* represents the multi-scale feature.

Finally, the dimensions of RGB_{msf} and $Depth_{msf}$ are mapped to a higher dimension C_1 through the Linear Projection operation to obtain the tokenized features: $RGB_{tokenized} \in \mathbb{R}^{C\times C_1}$ and $Depth_{tokenized} \in \mathbb{R}^{C\times C_1}$. The above steps can be formulated as:

$$RGB_{msf} = Concat(Flatten(msap(RGB_{in})), Flatten(mmsp(RGB_{in}))), \tag{7}$$

$$Depth_{msf} = Concat(Flatten(msap(Depth_{in})), Flatten(msmp(Depth_{in}))), \tag{8}$$

$$RGB_{tokenized} = LP(RGB_{msf}), \tag{9}$$

$$Depth_{tokenized} = LP(Depth_{msf}), \tag{10}$$

where subscripts *msmp* and *msap* represent multi-scale average pooling and multi-scale max pooling, respectively.

3.3.2. Channel attention correction

After converting the features of each channel into tokens through the I2T, we use the standard Transformer structure to model the features of all channels, and finally obtain the channel attention vectors of the two modals through MLP operation. Specifically, for the tokenized two modal features $RGB_{tokenized} \in \mathbb{R}^{C\times C_1}$ and $Depth_{tokenized} \in \mathbb{R}^{C\times C_1}$, first, a learnable modal-type encoding is added to the features of the two modals, so that it can distinguish different modal when performing self-attention calculations. Then, the features of the two modals are input into *N* Transformer Blocks, and the features $RGB_{cii} \in \mathbb{R}^{C\times C_1}$ and $Depth_{cii} \in \mathbb{R}^{C\times C_1}$ after channel information interaction are obtained, where subscript *cii* represents channel information interaction.

The calculation process of each Transformer Block can be formulated as:

$$z^l = CMMHSA(LN(z^{l-1})) + z^{l-1}, \tag{11}$$

$$z^l = MLP(LN(z^l)) + z^l, \tag{12}$$

where z^l represents the input of the l^{th} block, *LN* represents LayerNorm, and *CMMHSA* represents Cross-Modal Multi-Head Self-Attention.

After global modeling of all channel features through Transformer, we calculate the channel attention vectors $W_{rgb} \in \mathbb{R}^C$ and $W_{depth} \in \mathbb{R}^C$ of the two modals through MLP operation on the result of information interaction. Then we use the original feature maps $RGB_{in} \in \mathbb{R}^{H\times W\times C}$ and $Depth_{in} \in \mathbb{R}^{H\times W\times C}$ of the two modals to perform channel-wise multiplication with the channel attention vector. Finally, we perform a cross element-wise adding operation on the obtained results and the original feature map to obtain the corrected outputs: $RGB_{cor} \in \mathbb{R}^{H\times W\times C}$ and $Depth_{cor} \in \mathbb{R}^{H\times W\times C}$.The above steps can be formulated as:

$$RGB_{cii}, Depth_{cii} = Transformer(RGB_{tokenized} \oplus MTE_{rgb}, Depth_{tokenized} \oplus MTE_{depth}), \tag{13}$$

$$W_{rgb}, W_{depth} = MLP(RGB_{cii}, Depth_{cii}), \tag{14}$$

$$RGB_{cor} = RGB_{in} \oplus Depth_{in} \otimes W_{depth} \tag{15}$$

$$Depth_{cor} = Depth_{in} \oplus RGB_{in} \otimes W_{RGB}, \tag{16}$$

where $\oplus$represents element-wise adding, $\otimes$ represents channel-wise multiplication, and *MTE* represents modal-type encoding.

In the Transformer structure, due to the positive correlation between the computation of the MHSA part and the number of input tokens, the MHSA part will bring a significant computation when calculating in the spatial dimension. The MS-CAC module performs self-attention computation in the channel dimension (in the forward calculation process of the model, the number of channels is much smaller than the image resolution), to reduce its computation.

3.3.3. Workflow of MS-CAC module

In summary, the overall workflow of the MS-CAC module is as follows:

(1) Obtain multi-scale features of two modals by multi-scale pooling operations;
(2) Map the multi-scale features of two modals into high-dimensional space and add modal-type encoding;
(3) Global modeling of channel features for two modals using Transformer structure;
(4) Use MLP operation to calculate channel attention vectors for modeled features;
(5) Multiply the channel attention vectors of two modals by themselves;
(6) Finally, by cross element-wise adding operation, the channel corrected RGB features and depth features are obtained.

3.4. Global Feature Aggregation module

3.4.1. Principle of GFA Module

After performing channel attention correction on RGB feature and depth feature, we design the GFA module to exchange information in the spatial dimension, and fuse the features of the two modals into a single feature map for decoding. Compared with the traditional spatial

attention mechanism, the GFA module can capture the global features of an image and avoid the limitation of CNN on the size of receptive field. The detailed structure of the GFA module is shown in Fig. 4.

For the input features $RGB_{cor} \in \mathbb{R}^{H\times W\times C}$ and $Depth_{cor} \in \mathbb{R}^{H\times W\times C}$, since the GAF module performs cross-modal information interaction in the spatial dimension, it is necessary to embed the position information in the feature map. Inspired by CPVT [30], we introduce the position information through a depth-wise separable convolution with kernel size 3×3, stride 1×1, and padding size 1×1, and perform the element-wise adding operation with the input features through the residual connection to obtain features $RGB_{pe} \in \mathbb{R}^{H\times W\times C}$ and $Depth_{pe} \in \mathbb{R}^{H\times W\times C}$ with the positional information, where subscript *pe* represents the positional encoding. In addition to the positional encoding, we also add a learnable modal-type encoding.

The above steps can be formulated as:

$$RGB_{pme} = RGB_{cor} \oplus DWC_{3\times3}(RGB_{cor}) \oplus MTE_{rgb} \tag{17}$$

$$Depth_{pme} = Depth_{cor} \oplus DWC_{3\times3}(Depth_{cor}) \oplus MTE_{depth} \tag{18}$$

where subscript *pme* represents the positional encoding and modal-type encoding, and *DWC* represents the depth-wise convolution.

After obtain $RGB_{pme} \in \mathbb{R}^{H\times W\times C}$ and $Depth_{pme} \in \mathbb{R}^{H\times W\times C}$, we conduct information interaction in spatial dimension through the CM-MHSA module, and perform residual connection with the input. However, when calculate the self-attention in the spatial dimension, the computational complexity is related to the resolution of the input image, which will occupy a lot of computing resources and bring a lot of computational complexity. In the production workshop, all kinds of production information in the production process need to be obtained in real time. Therefore, we introduce the spatial reduction module proposed in PVT [27] to reduce the amount of computation through the sharing mechanism of *K* and *V*. The spatial reduction module first down-samples the feature map through reduction ratio R_i, and then calculates the *K* and *V* vectors on the down-sampled feature map, which can reduce the computation by R_i^2 times compared to the MHSA. Afterwards, the two modal features $RGB_{sii} \in \mathbb{R}^{H\times W\times C}$ and $Depth_{sii} \in \mathbb{R}^{H\times W\times C}$ after spatial information interaction are obtained through a LayerNorm layer, where subscript *sii* represents spatial information interaction. Finally, we fuse the features of the two modals into a single feature map through a simple 1×1 convolution. In addition, in semantic segmentation, local features are also needed to improve the robustness of semantic segmentation. So we use the original feature map to obtain local features through a depth-wise convolution $DWC_{3\times3}$, and fuse local features with global features through residual connections. Finally, the final output F_{out} is obtained through a batch normalization layer.

The above steps can be formulated as:

$$\begin{aligned} &RGB_{sii},\ Depth_{sii} = \\ &LN\big(CMMHSA\big(SR\big(RGB_{pme}, Depth_{pme}\big)\big) \oplus \big(RGB_{pme}, Depth_{pme}\big)\big), \end{aligned} \tag{19}$$

$$F_{global} = Conv_{1\times1}(Concat(RGB_{sii},\ Depth_{sii})), \tag{20}$$

$$F_{local} = CDRC(Concat(RGB_{cor}, Depth_{cor})), \tag{21}$$

$$F_{out} = BN\big(F_{local} \oplus F_{global}\big), \tag{22}$$

where *SR* represents the spatial reduction module, *LN* represents the LayerNorm, F_{global} represents the global features, *CDRC* represents a series of calculation processes of $Conv_{1\times1}$, $DWC_{3\times3}$, *Relu* and $Conv_{1\times1}$, F_{local} represents the local features, and *BN* represents the batch normalization.

3.4.2. Workflow of GFA Module

In summary, the overall workflow of the GFA module is as follows:

(1) Merge the features of the two modals according to channel dimensions, and add positional encoding and modal-type encoding;
(2) Reduce computational complexity by using spatial reduction to reduce the features of two modals;
(3) Interact information between the features of two modals in the spatial dimension through the CM-MHSA module;
(4) By adjusting the channel dimension of the feature map through 1×1 convolution, fuse it with the original input through residual linking.

Table 1
Detailed settings of the CMFormer series models.

	Output Size	Module Name	CMFormer-S	CMFormer-M	CMFormer-L
Stage1	$(\frac{W}{4},\frac{H}{4})$	MS-CAC	N_1=2	N_1=3	N_1=3
			H_1=1	H_1=1	H_1=1
			E_1=4	E_1=4	E_1=4
		GFA	R_1=8	R_1=8	R_1=8
			H_1=1	H_1=1	H_1=1
		Output Channel=64			
Stage2	$(\frac{W}{8},\frac{H}{8})$	MS-CAC	N_2=2	N_2=4	N_2=4
			H_2=2	H_2=2	H_2=2
			E_2=4	E_2=4	E_2=4
		GFA	R_2=4	R_2=4	R_2=4
			H_2=2	H_2=2	H_2=2
		Output Channel=128			
Stage3	$(\frac{W}{16},\frac{H}{16})$	MS-CAC	N_3=2	N_3=6	N_3=12
			H_3=5	H_3=5	H_3=5
			E_3=8	E_3=8	E_3=8
		GFA	R_3=2	R_3=2	R_3=2
			H_3=5	H_3=5	H_3=5
		Output Channel=320			
Stage4	$(\frac{W}{32},\frac{H}{32})$	MS-CAC	N_4=2	N_4=3	N_4=3
			H_4=8	H_4=8	H_4=8
			E_4=8	E_4=8	E_4=8
		GFA	R_4=1	R_4 = 1	R_4=1
			H_4=8	H_4=8	H_4=8
		Output Channel=512			
Encoder			pvtv2_b1	pvtv2_b2	pvtv2_b3
Decoder embedding dimension			128	256	512

3.5. Detailed settings of the CMFormer series

In order to provide more choices for different scenes, we further extend the CMFormer to a series models of different sizes, namely the CMFormer-Samll(CMFormer-S), the CMFormer-Medium(CMFormer-M),and the CMFormer-Large(CMFormer-L). To sum up, the hyperparameters of the CMFormer series models are shown in Table 1:

The hyperparameters are specified as follows:

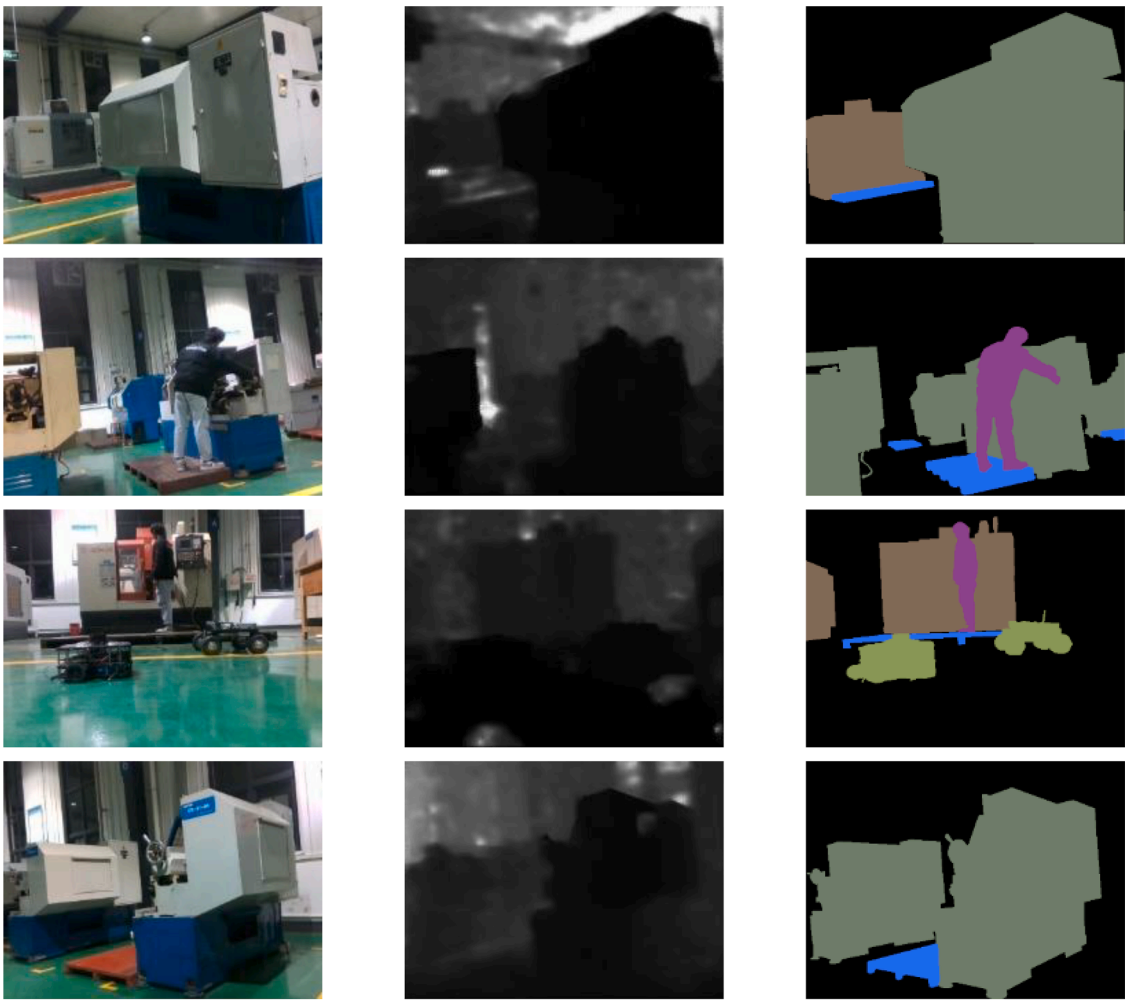
- N_i: the number of Transformer blocks in MS-CAC in Stage *i*;
- H_i: the number of heads per module in Stage *i*;
- E_i: the expansion ratio of the MLP layer in MS-CAC in Stage *i*;
- R_i: the reduction ratio of the SR in GFA in Stage *i*.

4. Experiments

In this section, we evaluate the performance of the CMFormer series models on the SOP dataset and NYU Depth v2 dataset(a mainstream RGB-D semantic segmentation public dataset).

4.1. Datasets and metrics

SOP [18]**:** The SOP dataset is a scene objects dataset for production workshop, including 2D/3D object detection, 2D instance segmentation, and 2D semantic segmentation. The 2D semantic segmentation contains 696 RGB-D images of the production workshop scene with a total of 7 semantic class labels, namely Person, Pedal, Robot, CNC milling machine (CNCMM), CNC lathe (CNCL), common milling machine (CMM) and common lathe (CL). The sample image resolution was 480 × 640. Examples of RGB images, depth image, and the semantic segmentation



(a)RGB images (b)Depth image (c)Ground Truth

Fig. 5. Examples of RGB images, depth image, and ground truth.

Table 2
Experimental environment.

Name	Version/Model
CPU	Intel Core i9 12900K
GPU	Nvidia RTX 3090
RAM	64GB 3600MHZ
Operating System	Ubuntu 18.04
Deep Learning Framework	Pytorch 1.8.2
Compute Unified Device Architecture	CUDA 11.1

Table 3
Specific details of hyperparameters.

Name	Value
Optimizer	AdamW
Learning Rate	0.00006
Learning Rate Schedule	Poly with power of 0.9
Warm-up Epochs	5
Batch Size	4
Epochs	100(on SOP dataset), 300(on NYU Depth v2 dataset)

ground truth are shown in Fig. 5. On this dataset, we use 500 instances for training, 96 instances for validation and 100 instances for testing.

NYU Depth v2 [19]**:** The dataset contains a total of 1449 RGB-D images of indoor scenes collected by Kinect, with a total of 40 semantic class labels. On this dataset, we use 795 instances for training and 654 instances for testing.

Metrics: We use two common metrics [4] in the field of semantic segmentation to evaluate our method, including Mean Pixel Accuracy (MPA) and Mean Intersection over Union(mIoU).

4.2. Implementation details

We use the Pytorch deep learning framework on Nvidia RTX 3090 GPU to build the CMFormer series models, and the detailed experimental environment is shown in Table 2.

For data augmentation, we use the same method in ACNet [9]. RGB image and depth image are randomly scaled, cropped, flipped and normalized. For RGB image, brightness and saturation are further enhanced. We take pvtv2 [41] pretrained on ImageNet [21] as the

Table 4
Comparison results with other models on SOP.

Models	Input	MPA(%)	mIoU(%)
FUSENet [39]	RGB-D	94.08	84.19
REDNet [5]	RGB-D	94.05	87.56
ACNet [9]	RGB-D	94.05	87.77
ESANet [43]	RGB-D	93.63	88.38
CMFormer-S	RGB-D	95.73	90.99
CMFormer-M	RGB-D	96.34	91.93
CMFormer-L	RGB-D	**96.74**	**92.98**

Table 5
Comparison results for all classes on SOP.

Models	Person	Pedal	Robot	CNCMM	CNCL	CMM	CL
FUSENet [39]	78.37	83.54	76.88	91.25	91.92	77.86	81.28
REDNet [5]	85.49	86.56	80.07	94.07	93.67	83.61	83.20
ACNet [9]	84.80	85.87	91.78	94.18	93.51	85.13	83.37
ESANet [43]	**87.94**	85.38	84.76	93.38	92.31	85.32	84.08
CMFormer-S	70.27	89.61	94.95	97.25	95.52	98.14	91.19
CMFormer-M	71.91	90.83	94.52	**97.78**	**97.18**	97.64	**93.63**
CMFormer-L	78.58	**92.74**	**95.68**	96.48	95.82	**98.41**	93.12

Table 6
Comparison results with state-of-the-art models on NYU Depth v2.

Methods	Models	Input	MPA (%)	mIoU (%)
CNN	Gupta et al. [38]	RGB-D	35.1	28.6
	Deng et al. [44]	RGB-D	-	31.5
	FCN [4]	RGB-D	46.1	34.0
	3DGNN [44]	RGB-D	55.7	43.1
	ACNet [9]	RGB-D	-	48.3
	RDFNet [6]	RGB-D	62.8	50.1
	ShapeConv [45]	RGB-HHA	63.5	51.3
	CANet [46]	RGB-D	64.6	51.5
	ESANet [43]	RGB-D	-	51.6
	NANet [11]	RGB-D	-	52.3
	SA-Gate [10]	RGB-HHA	-	52.4
Transformer	CMX(SegFormer-B2) [14]	RGB-HHA	-	54.2
	UCTNet(TR-Enc. TR-Enc.) [16]	RGB-D	-	55.3
	TransD-Fusion [13]	RGB-HHA	-	55.5
Ours	CMFormer-S	RGB-D	62.01	50.42
	CMFormer-M	RGB-D	66.36	54.12
	CMFormer-L	RGB-D	**68.00**	**55.75**

encoder and Semantic FPN [42] as the decoder. We refer to the hyperparameters setting used in CMX [14], because CMX [14] is the first method to apply Transformer to RGB-D semantic segmentation and has achieved good performance. The specific details of hyperparameter are shown in Table 3.

4.3. Experiment results on SOP dataset

The performances test of the CMFormer series models are compared with several open-source RGB-D semantic segmentation models on the SOP dataset, and the test results are shown in Table 4. Benefiting from the global attention mechanism of Transformer structure, the MS-CAC module and the GFA module proposed in this paper achieve more full feature correction and feature fusion by capturing the long-term dependency of the two modals, and fully mine the features of RGB image and depth image. Compared with ESANet [43], the CMFormer-S has improved MPA by 2.10% and mIoU by 2.61% on the SOP dataset [18]. With the increase of model size, the improvement of MPA and mIoU further expanded to 3.11% and 4.60%.

We further evaluate the segmentation performance of each class, the test mIoU indexes of the CMFormer series models and the other models are shown in Table 5. As can be seen from Table 5, except for the Person class, the other classes have a significant performance improvement compared with other methods. Due to the retention of multi-scale features, the performance gap between different objects with large scale differences (e.g., the difference between mIoU of Robot class and CNCMM class obtained from the CMFormer-L model is only 0.8%) has also been significantly improved. For the Person class, the mIoU is only 78.58%, we think that the number of Person class in the SOP dataset is smaller than that of other classes, while the Transformer structure cannot obtain better results when the amount of data is small. In addition, due to the different data distribution between ImageNet [21] dataset and SOP dataset, the weight of pre-training using ImageNet [21] dataset does not conform to the feature distribution in the production workshop. At the same time, there are many mechanical objects such as machine tools in the SOP dataset, which will suppress the features of the Person class in the feature extraction process, resulting in poor performance than CNN model.

4.4. Experiment results on NYU Depth v2 dataset

To verify the generalization of the CMFormer series models, we tested their segmentation performance on the NYU Depth v2 dataset, and compared it with the current state-of-the-art models, the test results are shown in Table 6. Among them, the MPA and mIoU of the CMFormer-M reached 66.36% and 54.12% respectively, surpassing all CNN-based methods in Table 6. Further, with the expansion of the model size, the CMFormer-L reached 68.00% MPA and 55.75% mIoU. It is close to the current excellent RGB-D semantic segmentation method based on Transformer. Compared with TransD-Fusion [13], although it uses Swin-B [31] with better performance as the backbone network, the CMFormer-L has better performance than TransD-Fusion [13].

For the experimental results of UCTNet [16] in Table 6, we exclude the influence of the uncertainty used in the paper on the results, and only compare the segmentation performance of the model itself. When the information flow restriction on uncertain nodes is added, the mIoU of UCTNet [16] can reach 57.6%, which is better than the method proposed in this paper.

For the experimental results of CMX [14], CMX [14] used a larger and better backbone network (SegFormer-B4, B5) [33] to further improve performance. However, due to the limitation of equipment conditions, we cannot train such a huge network. When the backbone network performance is close, CMX(SegFormer-B2) [14] and the CMFormer-M can achieve similar mIoU, while in CMX [14] with the SegFormer-B5 [33] backbone network, mIoU can reach 56.8%, which is also better than the method proposed in this paper.

Similarly, we analyze the segmentation performance of each class on NYU Depth v2 dataset, and the results are shown in Table 7. The CMFormer series models perform better than other methods in 33 classes (a total of 40 classes), especially in some smaller classes (e.g., box, bag, paper, clothes), which further prove that the proposed method retain the multi-scale features in the image as much as possible. In some classes (e.g., floor, ceiling, board), the CMFormer series models perform not well compared with other methods. We believe that on the one hand, the features of these objects are not obvious in the depth image and differ greatly from the appearance of the RGB image, on the other hand, in the feature correction part, the CMFormer series models only focus on the feature correction of the channel dimension and ignores the spatial dimension.

4.5. Ablation study

We conduct sufficient ablation studies on the NYU Depth v2 dataset to verify the effectiveness of the MS-CAC module and the GFA module. For a fair comparison, we remove all modules and use two parallel pvtv2

Table 7
Comparison results for all classes on NYU Depth v2.

Models	wall	floor	cabinet	bed	chair	sofa	table	door	window	bkshelf
FCN [4]	69.9	79.4	50.3	66.0	47.5	53.2	32.8	22.1	39.0	36.1
Gupta et al. [38]	68.0	81.3	44.9	65.0	47.9	29.9	20.3	32.6	39.0	18.1
Deng et al. [44]	65.6	79.2	51.9	66.7	41.0	55.7	36.5	20.3	33.2	32.6
RDFNet [6]	79.7	87.0	60.9	73.4	64.6	65.4	50.7	39.9	49.6	44.9
CANet [46]	**79.8**	**89.2**	65.6	72.8	64.4	65.7	47.6	47.6	49.1	44.3
CMFormer-S	70.95	88.09	62.84	71.86	64.88	61.44	49.95	39.04	48.11	46.61
CMFormer-M	72.55	88.94	**65.60**	73.89	68.98	68.30	51.30	42.87	49.76	47.44
CMFormer-L	73.71	89.02	65.02	**77.53**	**69.20**	**68.53**	**53.44**	**48.01**	**50.19**	**50.27**
Model	picture	counter	blind	desk	shelf	curtain	dresser	pillow	mirror	floormat
FCN [4]	50.5	54.2	45.8	11.9	8.6	32.5	31.0	37.5	22.4	13.6
Gupta et al. [38]	40.3	51.3	42.0	11.3	3.5	29.1	34.8	34.4	16.4	28.0
Deng et al. [44]	44.6	53.6	49.1	10.8	9.1	47.6	27.6	42.5	30.2	32.7
RDFNet [6]	61.2	67.1	63.9	28.6	14.2	59.7	49.0	49.9	54.3	39.4
CANet [46]	64.6	**72.3**	61.1	24.1	13.8	59.4	49.6	50.2	53.7	39.2
CMFormer-S	62.55	70.05	59.71	25.45	15.34	64.00	50.90	48.24	53.16	44.85
CMFormer-M	64.36	70.19	63.73	**29.30**	21.27	**70.01**	**57.32**	52.27	57.30	45.37
CMFormer-L	**66.56**	71.29	**64.51**	27.39	**22.40**	68.68	56.07	**54.30**	**60.53**	**46.27**
Model	clothes	ceiling	book	refridg	tv	paper	towel	shower	box	board
FCN [4]	18.3	59.1	27.3	27.0	41.9	15.9	26.1	14.1	6.5	12.9
Gupta et al. [38]	4.7	60.5	6.4	14.5	31.0	14.3	16.3	2.4	2.1	14.2
Deng et al. [44]	12.6	56.7	8.9	21.6	19.2	28.0	28.6	22.9	1.6	1.0
RDFNet [6]	26.9	69.1	35.0	58.9	63.8	34.1	41.6	38.5	11.6	54.0
CANet [46]	21.8	**77.8**	34.0	57.4	55.3	33.5	42.7	42.2	13.7	**73.1**
CMFormer-S	24.33	76.44	37.29	54.19	57.80	34.11	43.56	45.63	17.03	56.72
CMFormer-M	24.19	76.60	38.38	67.03	64.65	37.12	**48.86**	**49.86**	19.41	70.40
CMFormer-L	**28.55**	76.65	**39.63**	**70.20**	**69.86**	**40.02**	48.52	47.43	**21.52**	70.98
Model	person	nightstand	toilet	sink	lamp	bathtub	bag	othstr	othfurn	othprop
FCN [4]	57.6	30.1	61.3	44.8	32.1	39.2	4.8	15.2	7.7	30.0
Gupta et al. [38]	0.2	27.2	55.1	37.5	34.8	38.2	0.2	7.1	6.1	23.1
Deng et al. [44]	9.6	30.6	48.4	41.8	28.1	27.6	0	9.8	7.6	24.5
RDFNet [6]	80.0	45.3	65.7	62.1	47.1	57.3	19.1	30.7	20.6	39.0
CANet [46]	83.5	40.9	83.3	**68.1**	50.9	**63.2**	9.2	31.9	22.9	40.3
CMFormer-S	85.40	40.57	73.75	65.73	52.72	55.89	9.61	30.75	17.82	39.26
CMFormer-M	86.26	47.09	82.87	62.99	56.08	58.47	19.20	33.80	20.64	40.23
CMFormer-L	**87.79**	**54.29**	**83.71**	65.46	**56.17**	60.37	**23.59**	**34.88**	**25.23**	**42.39**

Table 8
Ablation for MS-CAC/GFA on NYU Depth V2.

MS-CAC	GFA	MPA(%)	mIoU(%)
baseline-S		58.65	47.33
√	×	61.20	49.39
×	√	59.74	48.36
√	√	**62.01**	**50.42**

Table 9
Ablation for both the MS-CAC module and GFA module on NYU Depth V2.

Models	Input	MPA(%)	mIoU(%)
baseline-M	RGB-D	63.69	51.74
CMFormer-M	RGB-D	**66.36**	**54.12**
baseline-L	RGB-D	64.53	52.85
CMFormer-L	RGB-D	**68.00**	**55.75**

[41] and Semantic FPN [42] as encoder and decoder for different modal. Then, the semantic predictions of the two modals are averaged to obtain the final prediction result. Through the above method, we obtained the baselines of three different sizes, namely baseline-Small(baseline-S), baseline-Medium(baseline-M), baseline-Large(baseline-L). During training, the same hyperparameters were used for each experiment.

For the CMFormer-S, when the MS-CAC module is ablated, the RGB feature and depth feature are extracted independently in their own branches; when the GFA module is ablated, we simply average the features of the two modals, and the results are shown in Table 8. Compared with the baseline-S, when only the GFA module is used, the MPA and mIoU reach 59.74% and 48.36%, increased by 1.09% and 1.03%; when only the MS-CAC module is used, the MPA and mIoU reach 61.20% and 49.39%, increased by 2.55% and 2.06%; when both modules are used together, the MPA and mIoU reach 62.01% and 50.42%, increased by 3.36% and 3.09%, which further proves the complementary ability of the MS-CAC module and the GFA module.

To demonstrate that the MS-CAC module and the GFA module are not constrained by the size of encoder and the decoder embedding dimensions, we conduct an overall ablation study on the two modules on the CMFormer-M and the CMFormer-L models, and the results are shown in Table 9. Compared with the baseline-M, the MPA and mIoU of the CMFormer-M reach 66.36% and 54.12%, increased by 2.67% and 2.38%; compared with the baseline-L, the MPA and mIoU of the CMFormer-L reach 68.00% and 55.75%, increased by 3.47% and 2.90%.

Table 10
Efficiency results. Parameters and FLOPs are estimated for inputs of RGB (480×640×3) and Depth (480×640×1).

Models	Params/M	FLOPs/G	FPS	mIoU(%)
CANet [46]	87.1	122.4	-	50.0
SA-Gate(ResNet50) [10]	63.4	204.9	-	50.4
CMFormer-S	**47.93**	**35.97**	43	50.42
CMFormer-M	81.75	68.38	27	54.12
CMFormer-L	131.41	128.92	18	**55.75**

4.6. *Efficiency and statistical analysis*

In the production workshop, because the equipment is running all the time, it is necessary to respond to emergencies in a timely manner to avoid losses. Therefore, the real-time performance of the model is particularly important. In Table 10, we show the number of parameters and computational complexity of our model, the CMFormer-S has only 47.93 million parameters and 35.97 GFLOPs of computation. Compared with SA-Gate [10], the CMFormer-S reduces the parameter amount of

Table 11
Results of five repeated experiments, the mean and standard deviation were calculated.

Datasets	Models	MPA(%)	mIoU(%)
SOP [18]	CMFormer-S	95.25±0.30	90.37±0.34
	CMFormer-M	96.06±0.37	91.50±0.44
	CMFormer-L	96.51±0.26	92.54±0.33
NYU Depth v2 [19]	CMFormer-S	61.79±0.28	49.99±0.22
	CMFormer-M	65.89±0.43	53.96±0.14
	CMFormer-L	67.53±0.47	55.25±0.35

the model by 24% and the computation amount by 82% with negligible difference in mIoU. We further test the FPS index of each model on our equipment(CPU: Intel i9 12900K, RAM: 64GB, GPU: RTX3090), and the FPS of the CMFormer-S can reach 43(far exceed the real-time performance index 1-10FPS of the workshop), which meets the real-time requirements in the production workshop scene. Although the CMFormer-M and the CMFormer-L models further improve mIoU, they have greater complexity. Therefore, the CMFormer-S is more suitable for the production workshop scene because of its high accuracy and real-time performance.

To avoid the impact of randomness caused by data enhancement and model parameter initialization on the results, we carried out five repeated experiments on each model to ensure the statistical significance of the results. The mean and standard deviation of five repeated experiments are shown in Table 11. It can be seen that the standard deviation of the model is within 0.5, the randomness of data enhancement and network parameter initialization will not have a significant impact.

4.7. Visualizations

To reflect the role of the MS-CAC module and the GFA module more intuitively, we visualized the feature maps of the first stage of the CMFormer-S on the NYU Depth v2 dataset, the results are shown in Fig. 6. We use the principal component analysis(PCA) method to reduce the dimension of the feature map to three channels for display. Among them, the yellow box represents the region where the RGB features are blurred but the depth features are clear; the red box represents the region where the RGB features are clear but the depth features are blurred. By applying the MS-CAC module and the GFA module, the features of two modals can be clearly displayed in the feature maps. For example, the backrest of the chair in the first row can be clearly displayed in the RGB image, while in the depth image, it cannot be clearly displayed due to the insufficient imaging distance of the depth camera. For the chair legs, it cannot be clearly displayed in the RGB image because of the dark light, but it can be clearly seen in the depth image. Through the MS-CAC module and the GFA module, we can clearly see in the feature map that whether it is the chair back or the chair legs, its features can be accurately extracted. In the second and third rows, the shutter and the picture are not prominent compared with the wall, and the depth camera cannot capture the slight bulge, so it is not obvious in the depth image. For the chair legs and the lamp post, RGB images cannot be clearly displayed due to the dark light. These features can also be clearly displayed in the feature map through the MS-CAC module and the GFA module. Similarly, for the TV set in the fourth row, its features cannot be clearly obtained in the depth image but can be accurately clearly in the feature map.

We have built a hardware testing platform and applied the

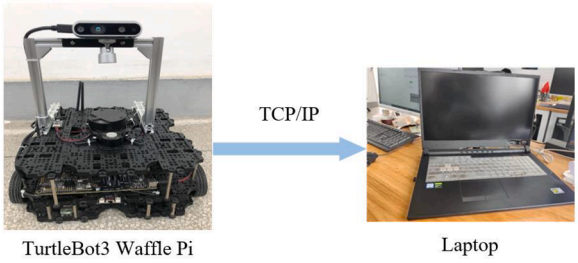


Fig. 7. Hardware testing platform.

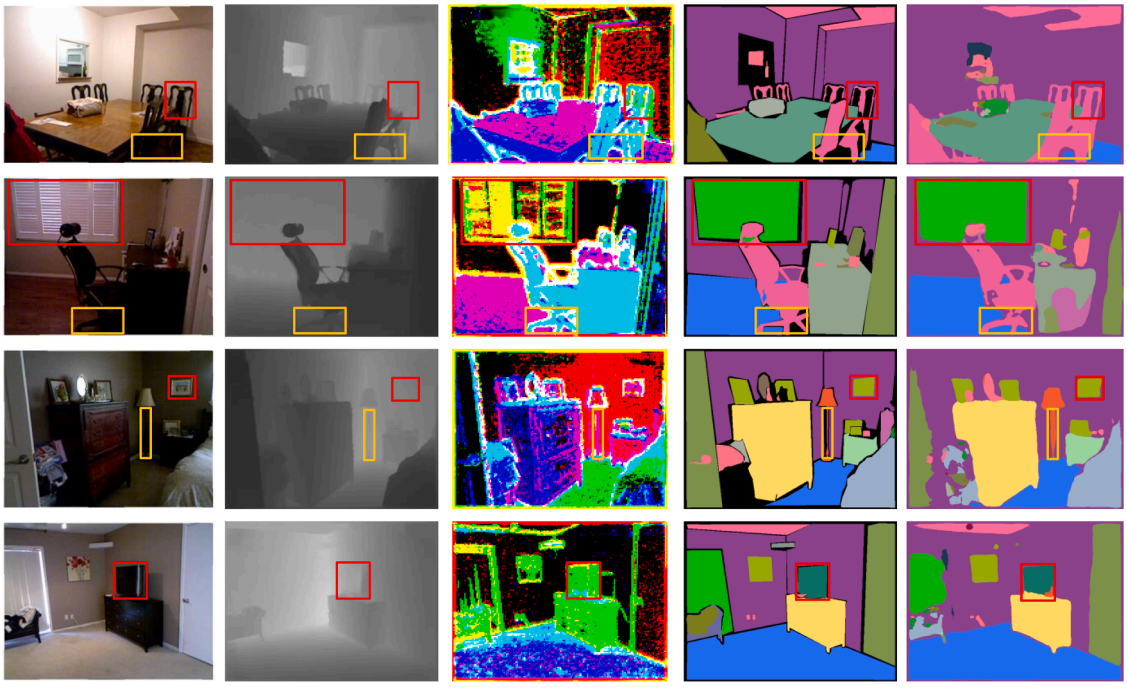


Fig. 6. Visualization results of the CMFormer-S on the NYU Depth v2 dataset. For each row, we show (a) RGB images, (b) Depth image, (c) feature maps after the MS-CAC module and the GFA module, (d) Ground Truth, (e) Results of Ours.

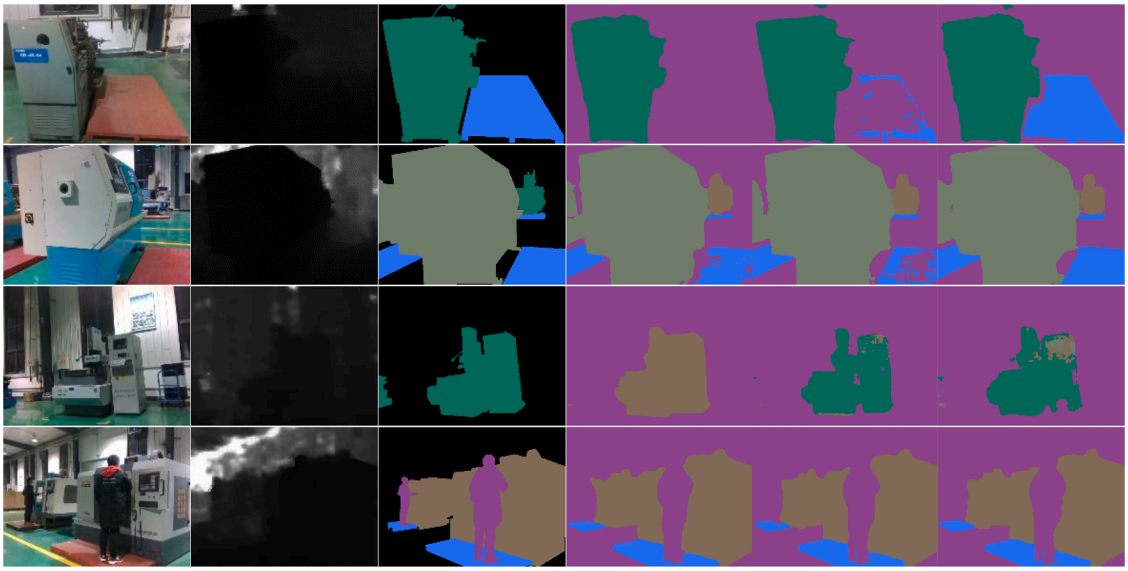


Fig. 8. Visualization results of three sizes of models on the SOP dataset. For each row, we show (a) RGB images, (b) Depth image, (c) Ground Truth, (d) Results of the CMFormer-S, (e) Results of the CMFormer-M, (f) Results of the CMFormer-L.

CMFormer in actual workshop scene, the hardware platform is shown in Fig. 7. Specifically, we used the TurtleBot3 Waffle Pi mobile robot equipped with the RealSense D435 RGB-D camera and Nvidia Jetson TX2 development board for image acquisition, and used a laptop as the computing platform to input the collected images into the model for inference.

We test the segmentation effects of the CMFormer series models, and the results are shown in Fig. 8. As can be seen from Fig. 8, as the model size increases, the object segmentation effect improves. For example, The CMFormer-S and the CMFormer-M in the first and second lines cannot completely segment the Pedal class in the picture, while the CMFormer-L can completely and accurately segment the Pedal class. In the third line, the CMFormer-S incorrectly classifies CL as CNCMM, and the CMFormer-M and the CMFormer-L are correctly classified. In the fourth line, as the model size increases, the outline of the Person class gradually becomes clearer.

5. Conclusion

Aiming at the multi-scale and real-time problems in the production workshop, this paper proposes a novel CMFormer, an RGB-D semantic segmentation model based on Transformer, which realizes the scene understanding in the production workshop. In order to better realize the information interaction between RGB image and depth image, this paper adopts the idea of Transformer and uses global attention to model the features between different modal. Specifically, we introduce the MS-CAC module and the GFA module to perform global modeling in both channel and space dimensions, and achieve accurate RGB-D semantic segmentation. At the same time, the method proposed in this paper greatly reduces the computational complexity of the model, which is more conducive to the application and deployment of production workshop scene.

To verify the effectiveness and generalization of the model, we conduct extensive comparative experiments and ablation experiments on both the SOP dataset and the NYU Depth v2 dataset. Experimental results show that the proposed CMFormer achieves state-of-the-art results, outperforming existing RGB-D semantic segmentation models. The experiment proves the advantages of the CMFormer in the production workshop scene, and expands the Transformer to the industrial manufacturing field, which reflects the potential industrial application value and provides theoretical reference for the research in this field.

Although the CMFormer model has good results in both datasets, there are still some limitations that require further research.

(1) For the sample imbalance problem of the dataset, this problem can be solved later by optimizing the loss function(e.g., Focal loss [15]).
(2) Due to Transformer's dependence on datasets, when there is less data in a certain class, the learning ability of the model is not as good as that of CNN model, which can be seen in Table 5. In the subsequent research, the problem of small amount of data can be solved through self-supervised learning(e.g., MoCo [47], MAE [48]).
(3) In actual workshop scene, it is possible to encounter the problem of camera shake causing poor imaging quality of depth image. In the subsequent research, this problem can be solved by improving the MHSA mechanism(e.g., UCTNet [16]) to suppress the uncertain information flow in the depth image.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgment

This paper was supported by the Sichuan Province Foreign and Overseas High-end Talent Introduction Program (No. 2022JDGD0013) and the Sichuan Province Science and Technology Support Program (No.2022YFG0198).

References

[1] Z. Wang, R. Song, P. Duan, X. Li, EFNet: enhancement-fusion network for semantic segmentation, Pattern Recognit. 9 (2021), 108023, https://doi.org/10.1016/j.patcog.2021.108023.

[2] T. Singha, D.S. Pham, A. Krishna, A real-time semantic segmentation model using iteratively shared features in multiple sub-encoders, Pattern Recognit. 140 (2023), 109557, https://doi.org/10.1016/j.patcog.2023.109557.
[3] C. Wang, C. Wang, W. Li, H. Wang, A brief survey on RGB-D semantic segmentation using deep learning, Displays 70 (2021), 102080, https://doi.org/10.1016/j.displa.2021.102080.
[4] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440, https://doi.org/10.1109/CVPR.2015.7298965.
[5] J. Jiang, L. Zheng,F. Luo, Z. Zhang, RedNet: Residual encoder-decoder network for indoor RGB-D semantic segmentation, arXiv preprint arXiv. (2018) 1806.01054. 10.48550/arXiv.1806.01054.
[6] S.J. Park, K.S. Hong, S. Lee, RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4980–4989, https://doi.org/10.1109/ICCV.2017.533.
[7] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141, https://doi.org/10.1109/TPAMI.2019.2913372.
[8] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19, https://doi.org/10.1007/978-3-030-01234-2_1.
[9] X. Hu, K. Yang, L. Fei, K. Wang, Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation, in: Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1440–1444, https://doi.org/10.1109/ICIP.2019.8803025.
[10] X. Chen, K.Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, G. Zeng, 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, Springer, Cham, 2020, pp. 561–577, https://doi.org/10.1007/978-3-030-58621-8_33.
[11] G. Zhang, J.H. Xue, P. Xie, S. Xie, S. Yang, Wang G, Non-local aggregation for RGB-D semantic segmentation, IEEE Signal Process. Lett. 28 (2021) 658–662, https://doi.org/10.1007/978-3-030-58621-8_33.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need. Advances in neural information processing systems. (2017) 30. 10.48550/arXiv.1706.03762.
[13] Z. WU, Z. ZHOU, G. Allibert, C. Stolz, C. Demonceaux, C. Ma, Transformer fusion for indoor Rgb-D semantic segmentation, Available at SSRN. (2022) 4251286. 10.2139/ssrn.4251286.
[14] H. Liu, J. Zhang, K. Yang, X. Hu, R. Stiefelhagen, CMX: Cross-Modal fusion for RGB-X semantic segmentation with transformers, arXiv preprint arXiv. 2203 (2022) 04838. 10.48550/arXiv.2203.04838.
[15] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 2020, pp. 318–327, https://doi.org/10.1109/TPAMI.2018.2858826.
[16] X. Ying, M.C. Chuah, UCTNet: uncertainty-aware cross-modal transformer network for indoor RGB-D semantic segmentation, in: Proceedings of the European Conference on Computer Vision, Springer, Cham, 2022, pp. 20–37, https://doi.org/10.48550/arXiv.1812.01243.
[17] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916, https://doi.org/10.1109/TPAMI.2015.2389824.
[18] Z. Tang, G. Chen, Y. Han, X. Liao, Q. Ru, Y. Wu, Bi-stage multi-modal 3D instance segmentation method for production workshop scene, Eng. Appl. Artif. Intell. 112 (2022), 104858, https://doi.org/10.48550/arXiv.1706.03762.
[19] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: Proceedings of the European Conference on Computer Vision, Berlin, Heidelberg, Springer, 2012, pp. 746–760, https://doi.org/10.1007/978-3-642-33715-4_54.
[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16×16 words: transformers for image recognition at scale, arXiv preprint arXiv. 2010 (2020) 11929. 10.48550/arXiv.2010.11929.
[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, Imagenet large scale visual recognition challeng, Int. J. Comput. Vision 115 (3) (2015) 211–252, https://doi.org/10.1007/s11263-015-0816-y.
[22] C. Sun, A. Shrivastava, S. Singh, A. Gupta, Revisiting unreasonable effectiveness of data in deep learning era, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 843–852, https://doi.org/10.1109/ICCV.2017.97.
[23] H. Touvron, M. Cord, M. Douze, A. Sablayrolles, Training data-efficient image transformers & distillation through attention, in: Proceedings of the International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357, https://doi.org/10.48550/arXiv.2012.12877.
[24] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F.E.H. Tay, J. Feng, S. Yan, Tokens-to-token vit: training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567, https://doi.org/10.1109/ICCV48922.2021.00060.
[25] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, Y. Wang, Transformer in transformer, Advances in Neural Information Processing Systems. (2021) 34. 10.48550/arXiv.2103.00112.
[26] C.F. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357–366, https://doi.org/10.1109/ICCV48922.2021.00041.
[27] W. Wang, E. Xie, X. Li, D-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: a versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578, https://doi.org/10.1109/ICCV48922.2021.00061.
[28] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, Cvt: Introducing convolutions to vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 22–31, https://doi.org/10.1109/ICCV48922.2021.00009.
[29] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, W. Wu, Incorporating convolution designs into visual transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 579–588, https://doi.org/10.1109/ICCV48922.2021.00062.
[30] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, C. Shen, Conditional positional encodings for vision transformers. arXiv preprint arXiv. 2102 (2021) 10882. 10.48550/arXiv.2102.10882.
[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022, https://doi.org/10.1109/ICCV48922.2021.00986.
[32] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. S. Torr, L. Zhang, Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890, https://doi.org/10.1109/CVPR46437.2021.00681.
[33] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: simple and efficient design for semantic segmentation with transformers, Advances in Neural Information Processing Systems, (2021) 34. 10.48550/arXiv.2105.15203.
[34] L. Yu, W. Xiang, J. Fang, Y.P.P. Chen, L. Chi, eX-ViT: a Novel explainable vision transformer for weakly supervised semantic segmentation, Pattern Recognit. 142 (2023), 109666, https://doi.org/10.1016/j.patcog.2023.109666.
[35] R. Strudel, R. Garcia, I. Laptev, C. Schmid, Segmenter: transformer for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7262–7272, https://doi.org/10.1109/ICCV48922.2021.00717.
[36] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the 2005 IEEE Computer Society Conference on computer Vision and Pattern Recognition (CVPR'05) 1, IEEE, 2005, pp. 886–893, https://doi.org/10.1109/CVPR.2005.177.
[37] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110, https://doi.org/10.1023/B:VISI.0000029664.99615.94.
[38] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from RGB-D images for object detection and segmentation, in: Proceedings of the European Conference on Computer Vision, Springer, Cham, 2014, pp. 345–360, https://doi.org/10.1007/978-3-319-10584-0_23.
[39] C. Hazirbas, L. Ma, C. Domokos, D. Cremers, Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture, in: Proceedings of the Asian Conference on Computer Vision, Springer, Cham, 2016, pp. 213–228, https://doi.org/10.1007/978-3-319-54181-5_14.
[40] Z. Shen, M. Zhang, H. Zhao, S. Yi, H. Li, Efficient attention: attention with linear complexities, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3531–3539, https://doi.org/10.1109/WACV48630.2021.00357.
[41] W. Wang, E. Xie, X. Li, D-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pvt v2: Improved baselines with pyramid vision transformer, Comput. Vis. Media 8 (3) (2022) 415–424, https://doi.org/10.1007/s41095-022-0274-8.
[42] A. Kirillov, R. Girshick, K. He, P. Dollár, Panoptic feature pyramid networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6399–6408, https://doi.org/10.1109/CVPR.2019.00656.
[43] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, H. Gross, Efficient rgb-d semantic segmentation for indoor scene analysis, in: Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2021, pp. 13525–13531, https://doi.org/10.1109/ICRA48506.2021.9561675.
[44] X. Qi, R. Liao, J. Jia, S. Fidler, R. Urtasun, 3d graph neural networks for rgbd semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5199–5208, https://doi.org/10.1109/ICCV.2017.556.
[45] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, Y. Li, ShapeConv: shape-aware convolutional layer for indoor RGB-D semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7088–7097, https://doi.org/10.1109/ICCV48922.2021.00700.
[46] H. Zhou, L. Qi, H. Huang, X. Yang, Z. Wan, X. Wen, CANet: Co-attention network for RGB-D semantic segmentation, Pattern Recognit. 124 (2022), 108468, https://doi.org/10.1016/j.patcog.2021.108468.
[47] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum Contrast for unsupervised visual representation learning, in: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), WA, USA, Seattle, 2020, pp. 9726–9735, https://doi.org/10.1109/CVPR42600.2020.00975.
[48] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 15979–15988, https://doi.org/10.1109/CVPR52688.2022.01553.

QINGJUN RU received his B.S. degree in Internet of Things Engineering from Zhengzhou University of Light Industry, China. Now he is a postgraduate student in Electronic

information from Chengdu University of Technology. His research interests focus on computer vision.

GUANGZHU CHEN received his Ph.D. degrees in Computer Engineering from Sichuan University, China. Now, he is a professor from Chengdu University of Technology. His research interests focus on computer vision, industrial data, internet security.

TINGYU ZUO received her B.S. degree in Network engineering from Chengdu Technological University, China. Now, she is a master of electronic information in Chengdu University of Technology Ω. Her research interests focus on computer vision.

XIAOJUAN LIAO received her Ph.D. degrees in Computer Engineering from University of Electronic Science and Technology of China. Now, she is an associate professor from Chengdu University of Technology. Her research interests focus on computer vision, industrial data.