Neural Computing and Applications (2023) 35:9627–9635
https://doi.org/10.1007/s00521-023-08199-4

ORIGINAL ARTICLE

TRCA-Net: stronger U structured network for human image segmentation

Li-Ying Hao[1] · Zhengkai Yang[1] · Yun-Peng Liu[1] · Chao Shen[2]

Received: 13 November 2021 / Accepted: 6 January 2023 / Published online: 31 January 2023

Abstract
Human image segmentation has been a practical and active research topic due to its wide range of potential application. There are some previous studies on manual, semi-automatic and automatic segmentation methods to investigate the semantic segmentation of human parts fully for real-world human analysis scenarios, but further research is still needed. This paper presents a novel semantic segmentation network, named TRCA-Net, for human image segmentation tasks. Having the TransUNet as the backbone, TRCA-Net incorporates Res2Net and Coordinate Attention to improve the performance. Res2Net blocks and Transformer can obtain better feature maps by encoding the input images. The Coordinate Attention in the decoder aggregates and upsamples the encoded feature maps, and connects to the high-resolution CNN feature maps for gaining accurate segmentation. The TRCA-Net can enhance finer details by recovering local spatial information. We compare the TRCA-Net with state-of-the-art (SOAT) semantic segmentation networks: the original U-Net, DeepLabv3+, and TransUNet. The experiment results have demonstrated that our proposed TRCA-Net outperforms these networks.

Keywords Semantic segmentation · Res2Net · TransUNet · Coordinate attention · TRCA-Net

1 Introduction

As one of the most fundamental and critical tasks in analyzing human in the wild, human parsing, or semantic segmentation has become a key enabling technology nowadays in a large number of application domains such as video surveillance [1], human behavior analysis [2], human part segmentation [3], medical image segmentation [4], and so on.

Semantic segmentation has recently witnessed great progress driven by the advancement of Convolutional Neural Networks (CNNs) [5], especially Fully Convolutional Networks (FCNs) [6]. Thanks to deeply learned features [7] and large-scale annotations [8], U-Net [9], a CNN-based Network, has become the state-of-the-art technology for human image segmentation.

Despite the excellent representational capabilities, the general limitation of CNN-based approaches goes to the incapability of displaying explicit remote relationship modeling due to the inherent limitations of convolutional operations [4]. As a result, these architectures often performs poorly for target structures that exhibit large differences in texture, shape, and size. To overcome this limitation, we employ the Res2Net [10] module together with Transformer [11] in this work. The Res2Net module builds hierarchical residual class links in a single residual block and fuses the features extracted by each hierarchical residual class link to improve the multi-scale capability of exploring CNN in a larger scope. As stated in [10], it essentially extends a new dimension (the number of feature groups in the Res2Net block), namely scale, which is an important and more effective factor in addition to the depth, width and cardinality dimensions.

By employing, dispense convolution operators entirely and relying on attention mechanisms solely, Transformers have emerged as alternative architectures to design for sequence-to-sequence prediction [11]. Unlike CNN-based approaches, Transformers are not only powerful in

✉ Chao Shen
shenchao@sce.carleton.ca

[1] College of Marine Electrical Engineering, Dalian Maritime University, Dalian, China

[2] Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada

modeling the global environment, but also exhibit excellent transfer capabilities for downstream tasks in the presence of large-scale pre-training. This success has been widely witnessed in the fields of machine translation and natural language processing [11, 12].

However, using transformer alone in the encoder process will lead to feature resolution loss [4]. Though the CNN-transformer works as a powerful tool to deal with the feature resolution loss, it can not better extract the features from the input image. To tackle this problem, in this work, we use the Res2Net-transformer for encoding, which not only eliminates the feature resolution loss but also improves the network layer range of the perceptual field.

To further improve the segmentation performance, we add Coordinate Attention [13] to the decoder process in the network. The Coordinate Attention decomposes the channel attention into two one-dimensional feature encoding processes, and collect features along two spatial directions, respectively. This allows a more aggregated feature map of the encoder. After being upsampled to recover the local spatial information, the aggregated feature maps are combined with the different high-resolution Res2Net features in the encoding path to achieve precise localization.

The main contribution of this paper is to put forward a semantic segmentation neural network with TransUNet as the backbone and Res2Net and Coordinate Attention. Compared the TRCA-Net with state-of-the-art (SOAT) semantic segmentation networks: the original U-Net, DeepLabv3+, and TransUNet, our proposed TRCA-Net has following performance advantages.

- Res2Net module is introduced in the feature extraction process to further improve the feature extraction ability.
- We use the Res2Net-transformer for encoding, which not only eliminates the feature resolution loss but also improves the network layer range of the perceptual field.
- We add coordinate attention to the decoder process in the network to further improve the segmentation performance.

The remainder of this paper is organized as follows: Sect. 2 provides related work; in Sect. 3, the proposed method and the baseline network are presented Sect. 4 gives the acquisition of input data, experimental setup and results. In Sect. 5, the conclusions of this study are shown.

2 Related work

2.1 Res2Net

Res2Net was first proposed in [10] as a simple yet effective module to explore the multi-system capabilities of CNNs over a larger range. It can be conveniently combined with existing state-of-the-art methods. Deng-Ping Fan et al. [14] proposed a new network for lung infection image segmentation. They used Res2Net as the backbone network for CT images to extract two sets of low-level features and three sets of high-level features. Using the powerful multi-scale capability of Res2Net, good CT image segmentation performance was achieved. In [15], a new residual multi-scale module with an attention mechanism drew on the multi-scale capability in Res2Net for single-image super resolution applications. Inspired by Res2Net, Yan Li et al. [16] developed the multiple temporal aggregation module, which divided spatiotemporal information and associated local convolution layers into a number of subsets.

2.2 Transformer

Transformer was first proposed in [11] for machine translation applications. Without reliance on CNNs, Alexey Dosovitskiy et al. [17] presented a Vision Transformer, which only applied a standard Transformer directly to sequences of image patches and completed image classification tasks very well in computer vision area. An improved Transformer architecture named Longformer was proposed in [18] to address the difficulty that computational requirements of self-attention increases guadratically with sequence length. By taking the advantage of the self-attention, Niki Parmar et al. [19] generalized the previous proposed Transformer of [11] to a sequence modeling formulation of image generation called Image Transformer. While Image Transformer can maintain much larger receptive fields per layer than traditional CNNs, it increases the size of images significantly.

2.3 Attention mechanisms

Initially created for machine translation [20], the Attention Mechanism has gradually grown in significance in the field of neural networks. An analogy to the human visual system can be used to explain attention mechanisms. The human visual system has a propensity to concentrate on specific details in an image that support judgment and dismiss irrelevant details. By inserting attention modules into CNN architectures, the performance of large-scale image classification tasks is improved substantially [21–23]. An effective module, called Bottleneck Attention Module (BAM) which was placed at each bottleneck of models, was proposed in [21]. By developing an inter-channel relationship, an attention module which is called as Squeeze-and-Excite (SE) was presented in [22]. In contrast to SE, Woo et al. [23] proposed the Convolutional Block Attention Module (CBAM) where both spatial and channel-wise attention was exploited. Despite these

improvements, only a few studies have used attentional mechanisms for image segmentation tasks.

3 Methods

The whole structure of our proposed TRCA-Net based on Res2Net [10], Transformers [11], and Coordinate Attention is shown in Fig. 1.

From Fig. 1, a convolutional layer is used to reduce the channel size of the reshaped features to the number of target classes in the last network layer, and then one can directly bilinearly upsample the feature maps to full resolution to predict the final segmentation results.

We can see that the decoder process together with the hybrid encoder forms a U-shaped architecture that allows feature aggregation at different resolution levels by skipping connections. The detailed architecture of the upsampling and intermediate skip-connection processes is shown in long dotted lines with arrows in Fig. 1.

Though combining CNN-Transformer [4] with combinatorial upsampling has achieved substantial performance, this strategy may not be the optimal choice for segmentation networks because the range of receptive fields in the encoder is not large and thus leads to a loss of low-level details. To gain a larger range of receptive fields, our TRCA-Net uses a hybrid Res2Net50-Transformer architecture as an encoder in the decoder process. Moreover, the upsampling and Coordinate Attention block includes an decoder, where the upsampling can expand the size of the feature map to the one of the input images and Coordinate Attention block can focus on the interested areas in the feature map.

The details of Res2Net, Transformer and Coordinate Attention are described below.

3.1 Res2Net

We first describe the encoder part of the TRCA-Net which is in the downsampling part. To combine the strengths of Res2Net block and Transformer, we use them to form the downsampling process.

The details of the Res2Net [10] block are shown in Fig. 2. After the input feature map is convolved using the 1×1 convolution, it splits the feature maps into s feature map subsets evenly. Except for X_1, each X_i has a corresponding 3×3 convolution, denoted by K_i. We denote y_i as the output of K_i. The feature subset X_i is added with the output of K_{i-1}, and is then fed into K_i. To reduce parameters the number of feature map subsets, we omit the 3×3 convolution for X_1. The 3×3 convolution operator has the potential to receive feature information from all feature segmentations. Each time a feature segmentation passes through the 3×3 convolution operator, the output may have a larger perceptual field. Different numbers and different combinations of receptive field sizes/scales are contained in the output of Res2Net module due to the effect of combinatorial explosion. In the Res2Net module, the segmentations are processed in a multi-scale manner, which facilitates the extraction of global and local information. We concatenate all the divisions and perform a single convolution on them in order to better combine the information at various scales. The splitting and tandem

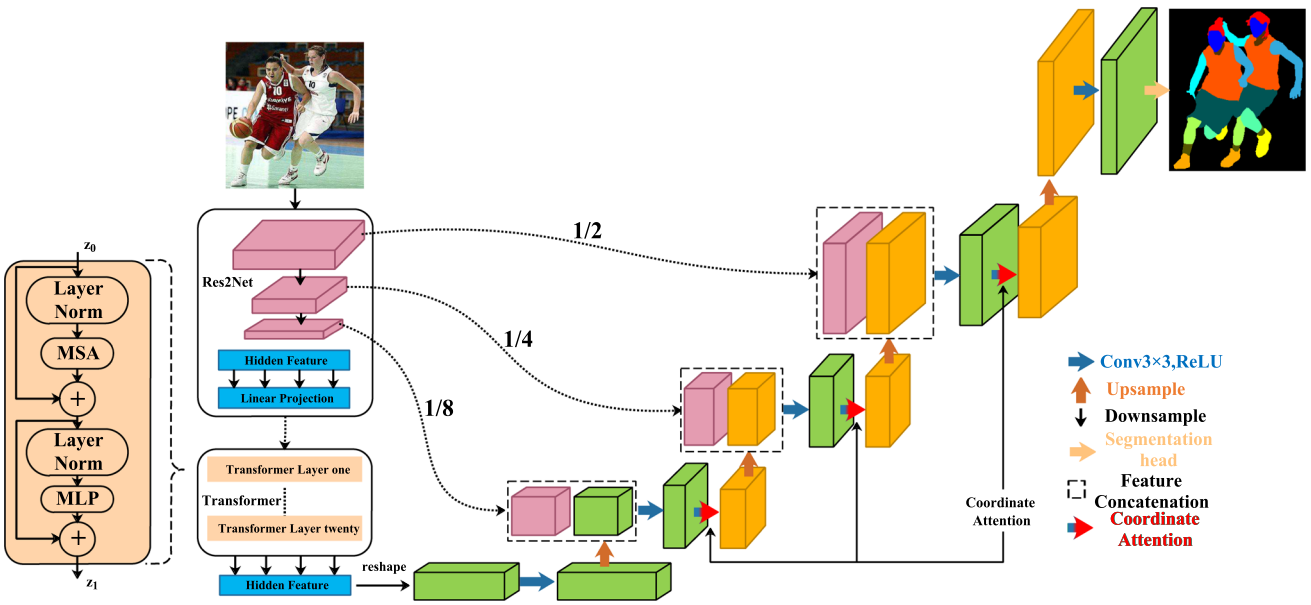

Fig. 1 Network architecture of TRCA-Net

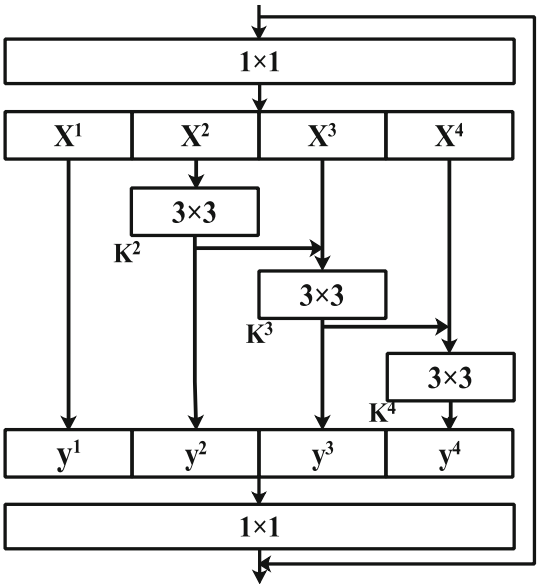

Fig. 2 The Res2Net block

approach can accelerate feature processing and force convolution.

3.2 Transformer

In the downsampling process, we perform tokenization by reshaping the output of the fortieth layer in the Res2Net50 [24] network into a sequence of flattened 2D patches. We map the vectorized patches into a latent D-dimensional embedding space using a trainable linear projection. To encode the patch spatial information, we learn specific position embeddings which are added to the patch embeddings to retain positional information, get the input of Transformer [11].

The Transformer encoder consists of L layers of Multihead Self-Attention (MSA) [25] and Multilayer Perceptron (MLP) [26] blocks. Therefore, the output of the ℓ-th ($\ell = 1, ..., 12$) layer can be written as follows:

$$z'_{\ell} = \text{MSA}\ (\text{LN}(z_{\ell-1})) + z_{\ell-1}, \tag{1}$$

$$z_{\ell} = \text{MLP}\ (\text{LN}(z'_{\ell})) + z'_{\ell}, \tag{2}$$

where $\text{LN}(\cdot)$ denotes the layer normalization operator , z_{ℓ} is the encoded image representation and z_0 is the product of merging all vectorized patches multiplied by the patch embedding projection plus the position embedding.

3.3 Coordinate attention

After reshaping the sequence of hidden features into the shape, we use a combined decoder that consists of three Coordinate Attention and multiple upsampling steps to reach full resolution from 3D feature map size to original image size, where each block is formed by a Coordinate Attention, two upsampling layers and one 3×3 convolutional layer. The Coordinate Attention structure is shown in Fig. 3.

Specifically, given the input feature map $M \in \mathbb{R}^{C\times H\times W}$, two spatially scoped pooling kernels $(H, 1)$ or $(1, W)$ are used to encode each channel along the horizontal and vertical coordinates, respectively. The above two transformations aggregate features along two spatial directions, respectively, to produce two direction-aware feature maps, which are combined. Then a shared 1×1 convolutional transformation function F_1 is performed on them, after a nonlinear activation function, the intermediate characteristic map of spatial information in the horizontal and vertical directions is obtained. The above process can be formulated as

$$f = \delta\left(F_1\left(\left[\frac{1}{W}\sum_{0\le i<W} m_c(h,i), \frac{1}{H}\sum_{0\le j<H} m_c(j,w)\right]\right)\right), \tag{3}$$

where $[\cdot,\cdot]$ is the concatenation operation along the spatial dimension, δ is a nonlinear activation function, $f \in \mathbb{R}^{C/r\times(H+W)}$ is the intermediate feature map, and r is the reduction rate used to control the block size. In order to reduce the complexity of the overhead model, in the literature [13], an appropriate reduction ratio r is used to reduce the number of channels of intermediate feature maps.

We then split f along the spatial dimension into two separate tensors $f^h \in \mathbb{R}^{C/r\times H}$ and $f^w \in \mathbb{R}^{C/r\times W}$. Another

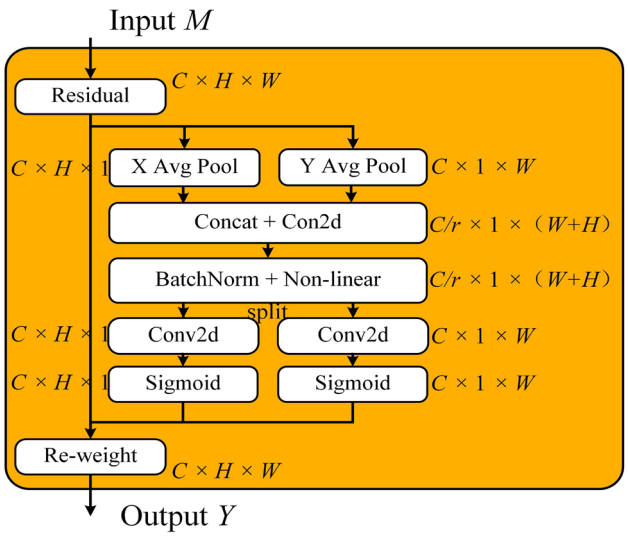

Fig. 3 The Coordinate Attention. “X Avg Pool” and “Y Avg Pool”, respectively, mean 1D horizontal global pooling and 1D vertical global pooling

two 1×1 convolutional transformations F_h and F_w are utilized to separately transform f^h and f^w to tensors with the same channel number to the input M, respectively, to pass the tensor through the sigmoid function. The two outputs are then expanded and used as attention weights, respectively. Finally, the output of our Coordinate Attention block Y can be written as

$$y_c(i,j) = m_c(i,j) \times \sigma(F_h(f^h)) \times \sigma((F_w(f^w)), \tag{4}$$

where σ is the sigmoid function. The above is the whole operation of Coordinate Attention. As described above, the attention along both the horizontal and vertical directions is simultaneously applied to the input tensor. Each element in the two attention maps reflects whether the object of interest exists in the corresponding row and column. This encoding process allows our Coordinate Attention to more accurately locate the exact position of the object of interest and hence helps the whole model to segmentation better.

In conclusion, the main mechanism that makes our network more suitable for human image segmentation tasks lies in that it combines the advantages of Res2Net and Coordinate Attention, which is stated in the following.

1. Due to the intricate outdoor environment, the structure and texture of the human body target are complex, and the size and shape are very different, so the network needs to have stronger feature extraction ability and accurate positioning ability of the extracted physical signs. The network structure proposed in this paper uses Res2Net to extract more abundant feature information.
2. We add coordinate attention in the decoder process. In this framework, it allows the encoder to map more converged features, locate the information of interest and suppress the useful information, so that the features can be accurately located.

4 Experiments and discussion

In this paper, we use the CIHP dataset [3] to verify the performance of the proposed TRCR-Net. In this dataset, all images are collected from real-world human activity scenes with 19 semantic labels and 28, 280 images are used as the training set.

For fair comparisons, we use one training batch and learning rate dynamic decreasing process for all models. We use nearest-neighbor interpolation to resize the original images to 512×512. The batch size is four images and the learning rate decreasing formula is as follows,

$$L = I_{\text{lr}}\left(1 - \frac{\text{iter}}{\text{max\_iter}}\right)^{\text{power}}, \tag{5}$$

where the initial learning rate I_{lr} is 0.001, real time training iteration *iter* is from 0 to 212100, maximum training iterations max_iter is 212100, the power is 0.9. We train the TRCA-Net at the same settings for 30 epochs. There are 28280 images for training in CIHP dataset. We believe that the number of samples is large enough, so the random data expansion method is not applied in this experiment. Our methods are implemented using the pytorch [27] framework. All networks are trained on a single graphics card NVIDIA GeForce GTX 3090 GPU.

4.1 Evaluation metrics

We evaluate all semantic segmentation models using Mean Pixel Accuracy (MPA) and Mean Intersection over Union (MIoU) criterion, which are statistical methods to test the similarity and diversity of the sample set. Intersection over Union IoU measures the similarity of a finite set of samples, which is defined as the size of the intersection set divided by the size of the concatenated sample sets. It is useful when the number of pixels in an image is unbalanced because the same weight for all classes. MIoU is the result of summing and averaging the ratio of the intersection of the predicted and true values for each category. PA is calculated as the ratio of the class pixel values of the predicted pairs to the total predicted values across all pixels. MPA is an extension of PA, which refers to the percentage of correctly classified pixels for each category and is calculated as the average of all PAs over all the classes. MPA and MIoU are calculated as evaluation metrics in Eqs. (6) and (7). In (6) and (7), k refers to the number of classes, and $k+1$ is the total number of classes including background. p_{ij} is the amount of pixels of class i inferred to belong to class j. In other words, p_{ii} represents the number of true positives, while p_{ij} and p_{ji} are usually interpreted as false positives and false negatives, respectively.

$$\text{MPA} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}}, \tag{6}$$

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{7}$$

Adopting the above evaluation metrics, we evaluate our proposed approach with CIHP dataset. The three baseline methods are the original U-Net [9], DeepLabv3+ [28] and TransUNet [4]. Table 1 summarizes the segmentation performances of each method. It can be seen that the proposed TRCA-Net outperforms the baseline methods in first three metrics such as MPA, MIoU and Recall.

Table 1 Comparison results of original U-Net, DeepLabv3+, TransUNet and proposed TRCA-Net

Method	MPA	MIoU	Recall	FLOPs(GLOPs)	Params(M)
U-Net [9]	21.00	17.58	83.64	40.669	4.321
DeepLabv3+ [28]	23.97	22.16	94.19	59.848	39.762
TransUNet [4]	22.65	19.53	86.73	134.988	93.252
TRCA-Net(ours)	**25.54**	**24.25**	**97.06**	**153.070**	**102.037**

The bold number denotes the results using the proposed algorithm, which is a common usage in the existing literature

As can be seen in Table 1, TRCA-Net outperforms the other methods. TRCA-Net increases 4.54, 1.57, and 2.89 points in MPA over original U-Net, DeepLabv3+, and TransUNet, respectively, and we also compare the performance of TRCA-Net, original U-Net, DeepLabv3+, and TransUNet in terms of MIoU. From the table we can see that TRCA-Net has the highest MIoU score, which is significantly better than other methods. The overall performance of TRCA-Net over the original U-Net, DeepLabv3+, and TransUNet in terms of recall is 83.63, 94.19, 86.73, and 97.06, respectively. Finally, we make the comparison with the above networks about floating-point operations per second (FLOPS) and parameters. Due to the introduction of the attention mechanism and Res2Net in the network, the computation and the number of parameters in the network are increased. Compared with TransUNet, which has the closest performance to TRCA-Net, the Params increased by 8.785M and FLOPS increased by 18.082G. Although the amount of computation is increased, it can be acceptable because the performance has achieved a great improvement.

Table 2 shows the results of each index of the four network models of each picture. It can be seen from Table 2 that among the three performance indicators of each picture, our model is better than those of other models. In addition to the quantitative results, we also

Table 2 Comparison results of original U-Net, DeepLabv3+, TransUNet and TRCA-Net in ten images

Name	Method	MPA	MIoU	Recall	Name	Method	MPA	MIoU	Recall
Image 1	U-Net [9]	23.78	19.60	85.93	Image 6	U-Net [9]	28.48	23.50	86.33
	DeepLabv3+ [28]	27.52	26.6	94.97		DeepLabv3+ [28]	30.50	28.15	94.89
	TransUNet [4]	22.20	18.32	84.79		TransUNet [4]	31.63	29.49	96.13
	TRCA-Net(our)	**27.66**	26.33	**96.31**		TRCA-Net(our)	**31.78**	**30.08**	**96.25**
Image 2	U-Net [9]	21.69	17.69	93.78	Image 7	U-Net [9]	20.41	18.94	89.46
	DeepLabv3+ [28]	22.39	20.62	97.44		DeepLabv3+ [28]	20.15	18.71	86.40
	TransUNet [4]	22.46	21.3	97.95		TransUNet [4]	21.45	19.74	89.56
	TRCA-Net(our)	**22.61**	**21.90**	**98.46**		TRCA-Net(our)	**23.15**	**22.30**	**97.03**
Image 3	U-Net [9]	15.21	11.67	73.70	Image 8	U-Net [9]	16.50	14.34	79.31
	DeepLabv3+ [28]	19.98	18.20	90.85		DeepLabv3+ [28]	22.53	20.96	94.10
	TransUNet [4]	18.25	16.57	84.79		TransUNet [4]	19.26	13.42	76.02
	TRCA-Net(our)	**23.97**	**21.96**	**97.97**		TRCA-Net(our)	**22.71**	**21.58**	**95.40**
Image 4	U-Net [9]	19.06	16.08	77.22	Image 9	U-Net [9]	24.00	19.70	70.80
	DeepLabv3+ [28]	22.37	20.27	94.47		DeepLabv3+ [28]	28.90	27.18	94.56
	TransUNet [4]	17.17	14.25	63.41		TransUNet [4]	26.21	20.22	87.61
	TRCA-Net(our)	**22.97**	**22.30**	**97.08**		TRCA-Net(our)	**30.53**	**28.74**	**96.78**
Image 5	U-Net [9]	19.31	15.39	84.39	Image 10	U-Net [9]	21.55	18.94	95.44
	DeepLabv3+ [28]	23.10	21.20	96.73		DeepLabv3+ [28]	22.24	19.79	96.71
	TransUNet [4]	25.89	23.51	90.71		TransUNet [4]	21.92	18.51	96.33
	TRCA-Net(our)	**27.94**	**26.43**	**97.94**		TRCA-Net(our)	22.21	**20.91**	**97.39**

The bold number denotes the results using the proposed algorithm, which is a common usage in the existing literature

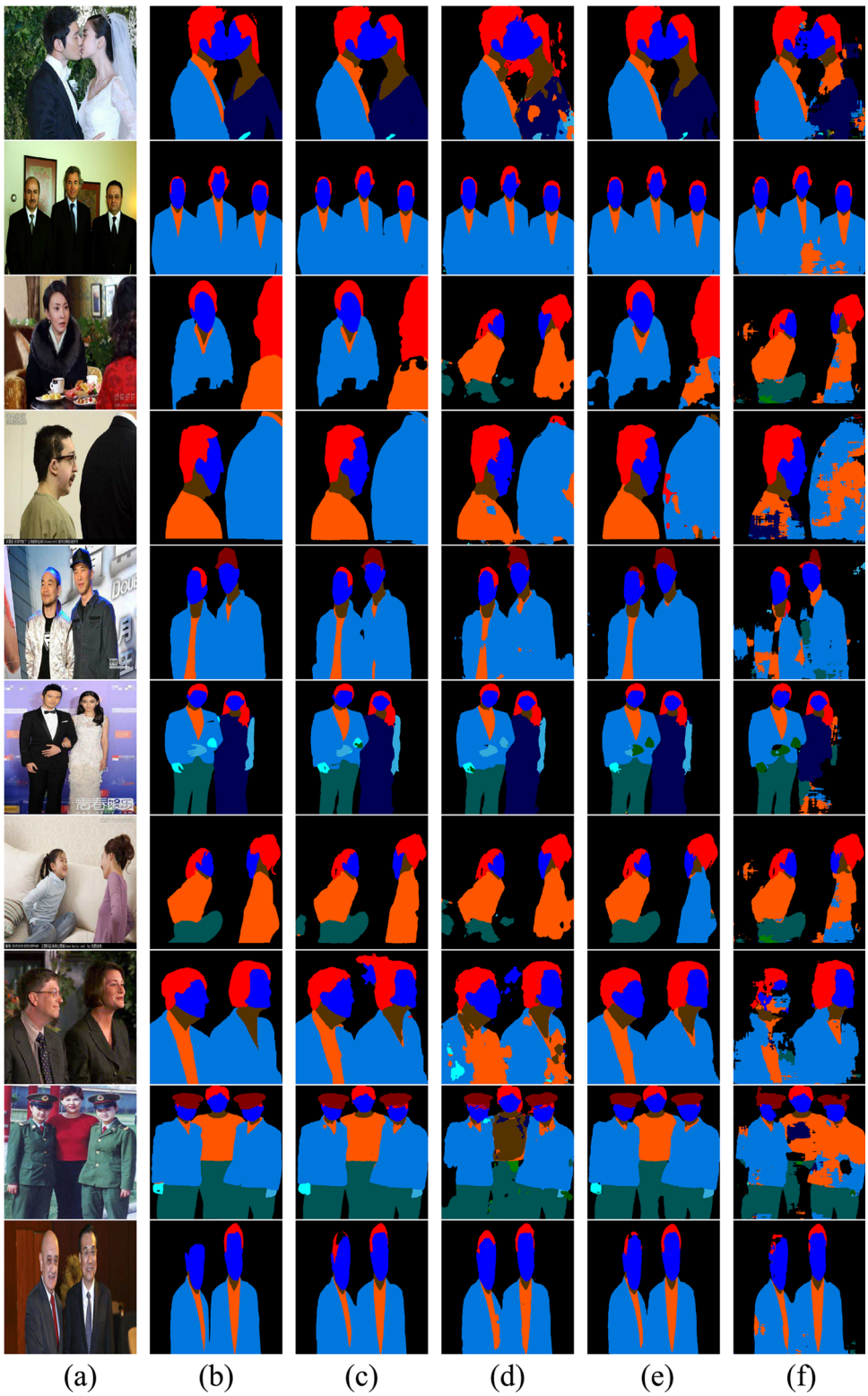

Fig. 4 Segmentation comparisons of TRCA-Net and other methods on ten images, **a** original image. **b** ground truth. **c** TRCA-Net(our). **d** TransUNet [4]. **e** DeepLabv3+ [28] and **f** U-Net [9]

provide qualitative results of segmented images using our method compared to the original U-Net, DeepLabv3+, and TransUNet, as shown in Fig. 4.

Figure. 4 shows the ground truth images using all methods and their segmentation results. From Fig. 4, the original U-Net gains the worst performance, especially in the images of eighth rows in the presence of ambient noise.

Table 3 Comparison results of TransUNet, TransUNet+Res2Net, TransUNet+CA and TRCA-Net

Method	Basics	Res2Net	CA	MPA	MIoU	Recall
TransUNet [4]	√	−−	−−	22.65	19.53	86.73
TransUNet [4]+CA [13]	√	√	−−	23.2	20.3	88.72
TransUNet [4]+Res2Net [10]	√	−−	√	24.4	23.7	95.58
TRCA-Net(ours)	√	√	√	**25.54**	**24.25**	**97.06**

The bold number denotes the results using the proposed algorithm, which is a common usage in the existing literature

It can be seen from the images in fourth and fifth rows that DeepLabv3+ and TransUNet can segment human images better than the original U-Net. However, the segmented images of DeepLabv3+ and TransUNet are not as accurate as the ones produced by TRCA-Net. We can see that similar segmentation results are obtained by TRCA-Net and DeepLabv3+ from Images 1, 2, 5, 8, 9, 10, and Images 3, 4, 6, 7 have shown that TRCA-Net has better segmentation results than DeepLabv3+.

4.2 Discussion

In this section, we verify our approach through two sets of comparison experiments. The experiments are all conducted with the same settings as subsection 4.1. All experiments for comparison experiments are conducted with the backbone of the TransUNet [4].

4.2.1 Ablation study

To verify the effectiveness of the proposed modules, we show the ablation study in this part. The performances of TRCA-Net have been compared with those of TransUNet, TransUNet+Res2Net and TransUNet+CA modules.

From Table 3, the MPA, MIoU, and Recall indices of TransUNet+CA and TransUNet+Res2Net are better compared with TransUNet. When Coordinate Attention is added, MPA, MIoU and Recall were increased by 0.55, 0.77 and 1.99, respectively. When Res2Net was added, MPA, MIoU and Recall were increased by 1.75, 4.17 and 8.85, respectively. The proposed TRCA-Net combines the advantages of Res2Net and Coordinated Attention and achieves better results, MPA, MIoU and Recall is increased by 2.89, 4.72 and 10.33, respectively.

4.2.2 Effectiveness of coordinate attention

In this part, we further discuss the effectiveness of Coordinate Attention used in our approach. We likewise perform an experiment to demonstrate the effectiveness of the Coordinate Attention for improving the feature representation. The performances of TransUNet+CA have been compared with those of TransUNet+SE and TransUNet+CBAM modules. It is worth noting that all three models have the same underlying structure except for the attention module and are trained using the same settings for fair comparisons. The experimental results for different attention modules are shown in Table 4.

Table 4 Comparison results of TransUNet+SE, TransUNet+CBAM and TransUNet+CA

Method	MPA	Mean IoU	Recall
TransUNet [4]+SE [22]	23.0	19.8	88.01
TransUNet [4]+CBAM [23]	22.8	19.5	88.14
TransUNet [4]+CA [13]	**23.2**	**20.3**	**88.72**

The bold number denotes the results using the proposed algorithm, which is a common usage in the existing literature

From Table 4, TransUNet+CA performs even better compared to TransUNet+SE and TransUNet+CBAM modules in terms of MIoU metrics with 0.5 and 0.8 points, respectively. The above experiments and data demonstrate the effectiveness of Coordinate Attention in the upsampling process.

5 Conclusion

In this paper, we have proposes a semantic segmentation network named TRCA-Net for human semantic segmentation. By combining the advantages of Res2Net, transformer and Coordinate Attention, TRCA-Net enhances the feature extraction ability and precise positioning ability of targets with complex structure and texture and large differences in size and shape. The neural network is more suitable for image segmentation tasks similar to outdoor human body segmentation.

To push the research boundary of outdoor human activity analysis for real-world scenes, we use a large-scale benchmark for an instance-level human parsing task, including 38, 280 pixelated annotated images with 19 semantic part labels. The experimental results on CIHP dataset show that our proposed method outperforms the SOTA segmentation methods, which implies the effectiveness and superiority of TRCA-Net.

Regarding to the advantages of the proposed network, it can be applied to mostly semantic segmentation and image

segmentation tasks such as face recognition detection, accurate location of face biometric features, and detection of different parts of human disease types in medical image segmentation. However, the computational complexity increased, and real-time capability is compromised because of the introduction of attention mechanism and Res2Net in the network. Due to the performance improvement, these losses are acceptable, and our future work will focus on addressing this issue.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 52171292, 51939001), Dalian Outstanding Young Talents Project (Grant No. 2022RJ05).

Data availability statements All data generated or analyzed during this study are included in this published article [3] and its supplementary information files.

References

1. Wang L, Ji X, Deng Q and Jia M (2015). Deformable part model based multiple pedestrian detection for video surveillance in crowded scenes. In VISAPP, pp.599-604. https://doi.org/10.5220/0004739105990604.
2. Gan C, Lin M, Yang Y, De Melo G and Hauptmann AG (2016). Concepts not alone: exploring pairwise relationships for zero-shot video activity recognition. In: AAAI, pp.3487-3493
3. Gong K, Liang X, Li Y, Chen Y, Yang M and Lin L (2018). Instance-level human parsing via part grouping network. In ECCV, pp.770-785. https://doi.org/10.1007/978-3-030-01225-0_47.
4. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al (2021). TransUNet: transformers make strong encoders for medical image segmentation. In CVPR. https://arxiv.org/abs/2102.04306
5. Hao L, Lie J, Guo G (2019) A multi-target corner pooling-based neural network for vehicle detection. Neural Comput Appl 32(18):14497–14506. https://doi.org/10.1007/s00521-019-04486-1
6. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(4):640–651. https://doi.org/10.1109/CVPR.2015.7298965
7. He K, Zhang X, Ren S, and Sun J (2016). Deep residual learning for image recognition. In CVPR, pp.770-778. https://doi.org/10.1109/CVPR.2016.90.
8. Gong K, Liang X, Zhang D, Shen X and Lin L (2017). Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing. In: CVPR, pp.932-940. https://arxiv.org/abs/1703.05446
9. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. Comput Sci. https://doi.org/10.1007/978-3-662-54345-0_3
10. Gao S, Cheng M, Zhao K, Zhang X, Yang M et al (2019) Res2Net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell 43(2):652–662. https://doi.org/10.1109/TPAMI.2019.2938758
11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al (2017). Attention is all you need. In: NIPS, pp 5998-6008
12. Devlin J, Chang MW, Lee K and Toutanova K (2018). Bert: pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805
13. Hou Q, Zhou D, Feng J (2021). Coordinate attention for efficient mobile network design. In: ECCV, pp.13713-13722. https://arxiv.org/abs/2103.02907
14. Fan DP, Zhou T, Ji GP, Zhou Y, Chen G, Fu HZ, Shen JB, Shao L (2020) Inf-Net: Auto-matic COVID-19 lung infection segmentation from CT S-cans. IEEE T Med Imaging 39(8):2626–2637. https://doi.org/10.1109/TMI.2020.2996645
15. Lan R, Sun L, Liu Z, Lu H, Luo X (2020) MADNet: a fast and lightweight network for single-image super resolution. IEEE T Cybern 51(3):1443–1453. https://doi.org/10.1109/TCYB.2020.2970104
16. Li Y, Ji B, Shi X, Zhang J, Kang B and Wang L (2020). Tea: Temporal excitationand aggregation for action recognition. In: CVPR, pp.909-918. https://doi.org/10.1109/CVPR42600.2020.00099.
17. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Houlsby N, et al (2020). An image is worth 16x16 words: transformers for image recognition at scale. In: CVPR. https://arxiv.org/abs/2010.11929
18. Beltagy I, Peters M E, and Cohan A (2020). Longformer: the long-document transformer. https://arxiv.org/abs/2004.05150v2
19. Parmar N, Vaswani A, Uszkoreit J, et al (2018). Image transformer. In:ICML, pp.4055-4064. https://arxiv.org/abs/1802.05751
20. Bahdanau D, Cho K and Bengio Y (2014). Neural machine translation by jointly learning to align and translate. In: RCLR. https://arxiv.org/abs/1409.0473
21. Park J, Woo S, Lee JY, Kweon IS (2018). Bam: bottleneck attention module. In: BMVC. https://arxiv.org/abs/1807.06514
22. Hu J, Shen L and Sun G (2018). Squeeze-and-Excitation networks. In: ECCV, pp.7132-7141. https://doi.org/10.1109/TPAMI.2019.2913372
23. Woo S, Park J, Lee JY, et al (2018). CBAM: convolutional block attention module. In: ECCV, pp.3-19. https://doi.org/10.1007/978-3-030-01234-2_1
24. F. Yu, D. Wang, E. Shelhamer, and T. Darrell (2018). Deep layer aggregation. In:CVPR, pp.2403-2412. https://doi.org/10.48550/arXiv.1707.06484
25. Jean-Baptiste Cordonnier, Andreas Loukas, Martin Jaggi (2020). On the relationship between self-attention and convolutional layers. In:ICLR. https://doi.org/10.48550/arXiv.1911.03584
26. Tolstikhin I O, Houlsby N, Kolesnikov A, et al.(2021) Mlp-mixer: an all-mlp architecture for vision[J]. Advances. IN:NIPS,pp.24261-24272. https://doi.org/10.48550/arXiv.2105.01601
27. Paszke A, Gross S, Massa F, et al (2019). Pytorch: An imperative style, high-performance deep learning library[J]. In: NIPS,pp.3-12. https://doi.org/10.48550/arXiv.1912.01703
28. Liang CC, Yukun Z, George P, Florian S et al (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV, pp.8-14. https://doi.org/10.1007/978-3-030-01234-2_49.