A deep learning algorithm for KOL segmentation on social media videos

Cheng Yang

ych2tj@gmail.com

Auckland University of Technology

Fucheng Zheng

Auckland University of Technology

Duaa Zuhair Al-Hamid

Auckland University of Technology

Peter han Joo Chong

Auckland University of Technology

Patrick Lam

Zyetric Technologies Ltd

Research Article

Keywords: deep learning, image segmentation, social media video, convolutional neural networks

Posted Date: January 15th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3851659/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at International Journal of Pattern Recognition and Artificial Intelligence on November 25th, 2024. See the published version at https://doi.org/10.1142/S0218001424520281.

A deep learning algorithm for KOL segmentation on social media videos

Cheng Yang[1], Fucheng Zheng[1], Duaa Zuhair Al-Hamid[1], Peter Han Joo Chong[1], Patrick Lam[2]

[1]Auckland University of Technology, Dept of Electrical and Electronic Engineering, 55 Wellesley Street East, Auckland Central, New Zealand.
[2]Zyetric Technologies Ltd, Unit 337, 3/F, Building 19W, Hong Kong Science Park, Shatin, N.T., Hong Kong, China.
Corresponding author: Cheng Yang; ych2tj@gmail.com
ORCID: Cheng Yang: 0000-0002-4654-5861
Peter Han Joo Chong: 0000-0002-5375-8961
Duaa Zuhair Al-Hamid: 0000-0002-2832-7190

Abstract
Nowadays, there is high commercial demand in advertising for KOL (Key Opinion Leader) to advertise commercial products in their social media videos. One effective technique is the product replacement which places the products virtually in the videos. However, one of the challenges of placing the products virtually is the KOL segmentation. Because KOLs often hold products in front of them, which requires the segmentation to segment not only human but also different products. This paper introduces the state-of-the-art deep learning method, namely RSUDISNet, for KOL segmentation on social media video. The proposed deep convolutional neural network (CNN) can segment both KOL and different products which block the KOLs. The proposed technique integrates two CNN technologies. One is the matting objective decomposition network (MODNet), which segments KOLs well but not the products blocking the KOLs. The other one is the two-level nested U-structure network (U2Net) based on salient object detection method to segment the objects well, but not the KOL. The key technique of U2Net is the residual U-block (RSU), which can build neural network architecture deeper, while saving computing. This research employs the RSU block to embed the MODNet to overcome the problem of KOL segmentation from U2Net. Since both MODNet and U2Net are lightweights, the combined network can be used for real-time scenario. After that, the intermediate supervision (IS) training strategy is utilized to overcome the overfitting, which increases the accuracy to a higher level. The experimental results show that our proposed method outperforms the MODNet and U2Net.

Keywords deep learning; image segmentation; social media video; convolutional neural networks

1 Introduction

Artificial neural network (ANN) plays an important role in the industry advertising. One of the ANN applications is the product replacement in the KOL (Key Opinion Leader) social media videos. The KOLs normally compare or evaluate different commercial products in their videos. The product replacement technique can be used to add the advertisement (AD) image behind the KOL in a video such that it does not distract the audiences from watching the video. This is aided by foreground segmentation using convolutional neural networks (CNN). Such KOL videos offer a product substitution with a sizable potential market. As a result, KOL segmentation is the key challenge to make a considerable impact on the product replacement market.

Although, on social media video, KOL segmentation is comparable to human segmentation, there are some factors to make KOL segmentation in video product replacement more difficult. First, a KOL may be holding a variety of items, including gadgets like fragrances, cell phones and food mixers or food like lobster and beef. They should be segmented from the KOL body as well. Second, the KOL can be in close proximity to the camera, allowing for clear displays of their fingers and hair. These factors do require highly precise segmentation in the fine details.

Human segmentation is the initial concept for KOL segmentation. When it comes to human segmentation, deep learning techniques can be more precise than conventional computer vision techniques [1]. A deep neural network was utilized by [2] to separate the individual from the background. The deep learning algorithm has the advantage of not being colour, pattern, or shape sensitive. Previous studies have investigated the human body segmentation [3-5]. A backbone-branches architecture neural network in [3] segments human body parts, like face, hair, arms, hands, chest and legs. The method first learnt human poses, and then used the pose results to improve segmentation performance. The detail of the human figure is difficult to be segmented using traditional

methods. Therefore, human matting technology has emerged. To fulfil the human instance matting challenge, the authors in [4] proposed a multi-instance refinement neural network architecture. The human matting generated an alpha matte that reveals more detail of the human figure, particularly the hair. Another human portrait matting method which was reported by [5] utilized the objective decomposition network. This method was lightweight, yet it produced human matting with high accuracy. Human segmentation and matting can be performed accurately by the deep learning algorithm. However, the KOL segmentation may be affected by other objects when a KOL is holding the objects in front of him/her.

Various objects can be segmented using a salient object detection approach. To segment salient items, the study in [6] developed a CNN-based method. The CNN includes a global guiding module to determine the potential location of the salient items. The details of the conspicuous items were then refined using a feature aggregation tool. Ke and Tsubono [7] designed a CNN with a contour saliency blending module. This module extracted the contour features of the salient object, which could improve the salient object detection. Qin et al in [8] introduced a salient object detection neural network with a significantly complex architecture. Their CNN included ReSidual U-blocks, which boosted the depth of the neural network without increasing the computational cost. The salient object detection algorithm can segment salient objects from images. However, it cannot be used directly in KOL segmentation. If a KOL holds an object in front of his/her face, the salient object detection algorithm may recognise the object but disregard the KOL body. As a result, the object is segmented as foreground. Conversely, the KOL body is segmented into background regions.

This paper focuses on designing a CNN by combining human segmentation and salient object detection techniques. The related CNN used for human segmentation is the matting objective decomposition network (MODNet) [5]. This neural network is lightweight. It includes a low-resolution module, a high-resolution module and a fusion module. The low-resolution module extracts the global features. This module can accurately find the potential location of the target object. However, the high-resolution branch is not very accurate for the detail segmentation. To increase the accuracy, the ReSidual U-block from U2Net [8] is employed to build deeper high-resolution branch and fusion branch. This leads to the MODNet going deeper. However, it does not increase the computational cost. Therefore, the proposed new neural network, **RSUDISNet**, combining MODNet and U2Net can outperform the MODNet and U2Net alone. Finally, to solve the overfitting problem of the deeper architecture, the intermediate supervision (IS) [9] training strategy is utilized to optimize the training of the neural network.

2 Related works

This paper aims to segment KOLs using social media videos. As human segmentation is a study strategy that is comparable, the latest technology used in human segmentation is the deep learning algorithm. Herein, the authors in [2] introduced a deep neural network (DNN) to segment human from the surveillance video. The architecture of the neural network included encoder-decoder CNN. The encoder was similar to VGG-16 [10] CNN, which is used to extract features. The decoder included a set of operations of deconvolution and up-sampling to reach the input image size. This CNN segmented moving human as foreground from the background. The limitation of the deep learning algorithm was the big data requirement. Making synthetic data is one method of expanding the data set without identifying it. However, Lin et al [3] reported that the synthetic data training had worse results compared to the real-world data and manual labelling. This is mainly due to the domain gap between the synthetic and real-world domains. They found that the human skeleton or pose could effectively bridge the domain gap. Their CNN consisted of two main components. The first module learned about body parts and human in the synthetic domain. The second module shared parameters with the first component for the real-world data input training. When deep learning for human segmentation was getting popular, there was research to analyse human parsing by using human parts segmentation. Luo et al [11] proposed the Macro-Micro Adversarial Net (MMAN) to segment human parts, such as the head, torso, upper arms, lower arms, upper legs and lower legs. Their network has two discriminators. One discriminator (Macro D) acted on the low-resolution label map and avoided the misplaced body parts. The other discriminator (Micro D) focused on the high-resolution label map to address local inconsistency, like blur and hole. This network enforced local and semantic consistency explicitly, and also handled the high-resolution images. Another example is [12], who presented Parsing R-CNN for instance-level human analysis. Their network included FPN backbone, RPN network, RoI Align to extract human features. The bounding box branch detected each human. In parallel, the parsing branch

segmented human body parts. This network did human parsing on multiple human bodies on one image. Fang et al [13] utilized Weakly and Semi-supervised learning to do the human parsing. Their key idea is to mine the anatomical similarity among human to transfer the parsing results of a person to another person with a similar pose. The results can outperform some supervised learning methods.

In addition to human segmentation, human matting technology further refines human segmentation. This technique distinguishes between the background and the foreground. However, due to human hair's translucent and web-like appearance, some areas might not be known, which are also detected by matting method. For human matting, there are numerous deep learning methods. In addition to an RGB image and an alpha matte, the early human matting approach required a tri-map in order to train a CNN [14]. Currently, although many technologies attempt to reduce the use of tri-maps, they are nonetheless beneficial [15]. Sun et al [16] introduced a method to obtain better alpha mattes by incorporating into their framework the semantic classification of matting regions. They extend the conventional tri-map to a semantic tri-map. Meanwhile, their CNN had a multi-class discriminator to regularize the alpha prediction at the semantic level, and content-sensitive weights to balance different regulation losses. Wu et al [17] built a CNN which included three networks: pose network, tri-map network and matting network. They also appended a tri-map refinement module and utilized gradient loss to provide a sharper alpha matte. Unless the normal idea which put a tri-map into the input of a CNN, the ground truth tri-map is used to compare the output of the tri-map network. Therefore, the test of the neural network didn't require a tri-map for input. The limitation of tri-map is that it costs the manual labelling work. A method from [18] tried to reduce the labelling work by using coarse annotation. They proposed to use coarse annotated human data coupled with alpha matte to boost end-to-end semantic human matting without tri-maps as extra inputs. Their CNN also has three parts: the mask prediction network estimated the coarse semantic mask; a quality unification network to unify the quality of the coarse mask outputs; and finally, a matting refinement network took the unified mask and the input image to predict the final alpha matte. Some human matting approaches could train a CNN without manual labelling of the tri-map. These are tri-map-free networks. Chen et al [19] developed a Semantic Guided Human Matting (SGHM) which was built on a semantic human segmentation network. An image was first down-sampled for the shared encoder, then the segmentation decoder was used to generate a coarse semantic mask prediction. This decoder also shared features with the matting decoder, which refined the human margin and predicted the alpha matte. A video human matting method [20] utilized a recurrent architecture to exploit temporal information in videos to improve matting quality. The architecture included an encoder to extract features followed by three up-sampling blocks of the recurrent decoder. After that, a deep guided filter module was used for high-resolution (4K) prediction. Another tri-map free human matting method is from [5] which was called a matting objective decomposition network (MODNet). The MODNet included three branches. The Low-Resolution branch was to locate the human region in the input image. The High-Resolution branch extracted the features for detail part prediction. The final fusion branch predicted the alpha matte. This was also a light-weight architecture which run 67 frames per second.

Human segmentation and matting can segment human well. However, KOLs on social media also hold different objects which are not easy to be segmented by the human segmentation method. The salient object detection (SOD) method is good at segmenting the foreground objects. The popular type of the CNN is the context-aware or local-global architecture [6, 21-23]. A good example is [21] which proposed a global-local collaborative architecture, which included a global (GCM) and a local (LCM) correspondence modelling to extract comprehensive inter-image corresponding relationships among different images from the global and local perspectives. The inter-image relationships of the GCM and LCM are integrated through a global-local correspondence aggregation (GLA) module. Finally, the intra and inter-features were adaptively integrated through an intra-inter weighting fusion (AEWF) module to learn co-saliency features and predict the saliency map. In another example, Liu et al[6] utilized pooling techniques for salient object detection. A global guidance module (GGM) was to guide the location information of the potential salient object at different feature levels. A feature aggregation module (FAM) seamlessly fused the coarse-level semantic information with the fine-level features. Finally, these two modules were utilized to refine the high-level semantic features and obtain detail enriched saliency maps.

Another common type of CNN for salient object detection is the boundary or contour-aware architecture [7, 24, 25]. A special example is [7] which designed a contour-saliency blending module to exchange information between contour and

saliency. They utilized recursive CNN to increase contour-saliency fusion while keeping the total trainable parameters the same. Furthermore, a stage-wise feature extraction module was designed to help the network pick up the most helpful features from previous intermediate saliency predictions. In addition, [24] introduced a boundary-aware salient object detection (BASNet). The architecture was composed of a densely supervised encoder-decoder network and a residual refinement module. A hybrid loss was designed to guide the network to learn the transformation between the image and the ground truth by fusing binary cross-entropy, structural similarity and intersection-over-union losses. The predict-refine architecture and the hybrid loss was able to segment the salient regions and predicted the fine structures with clear boundaries.

Furthermore, some salient object detection networks are designed to have efficient architecture, which is lightweight and fast speed [8, 26-28]. A famous theory is a two-level nested U-structure network (U2Net) [8]. The network was designed to capture more contextual information from different scales by using the mixture of receptive fields of different sizes in the proposed ReSidual U-(RSU) blocks. It also increases the depth of the whole architecture without significantly increasing the computational cost. This network enabled the training without using backbones from the image classification tasks. The running speed was 30 FPS on GTX 1080Ti GPU with 172MB weight.

3 The Methodology

The proposed RSUDISNet is inspired by MODNet[5], U2Net[8] and IS[9]. MODNet is designed for human matting. The architecture includes a low-resolution branch, a high-resolution branch and a fusion branch. MODNet can segment humans based on position and boundary details thanks to these three branches. The U2Net has RSU blocks allow for considerably deeper network design without the need for a big backbone network. The IS calculate the training loss in high dimensions which significantly overcomes the overfitting.

3.1 The RSU block

In this paper, the MODNet is made deeper by using the ReSidual U-block (RSU) [8]. The RSU block captures intra-stage multi-scale features. This architecture makes the neural network go deeper but saves computational consumption. The architecture of RSU block is shown in **Fig. 1**. In the figure, L is the number of layers in the encoder. The C_{in} and C_{out} denote input and output channels, and M represents the number of channels in the internal layers. The RSU mainly includes three components:

1) An input convolution layer transforms the input feature map $\boldsymbol{x}$ ($H \times W \times C_{in}$) to an intermediate map with the channel of C_{out}. This is for local feature extraction.
2) A U-structure-like symmetric encoder-decoder architecture with a height of L takes the intermediate feature map as input and leans to extract and encode the multi-scale contextual information. The larger L leads to a deeper residual U-block, more pooling operations, a larger range of receptive fields and richer local and global features. The extraction of multi-scale features from input feature maps is controlled by this parameter with arbitrary spatial resolutions. The multi-scale features are extracted from gradual downsampling and encoded into high resolution by progressive upsampling, concatenation and convolution. This process

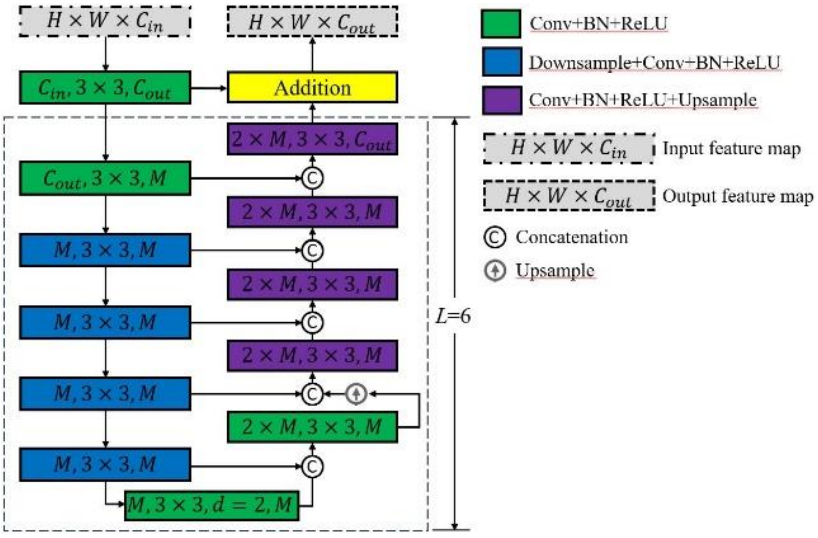
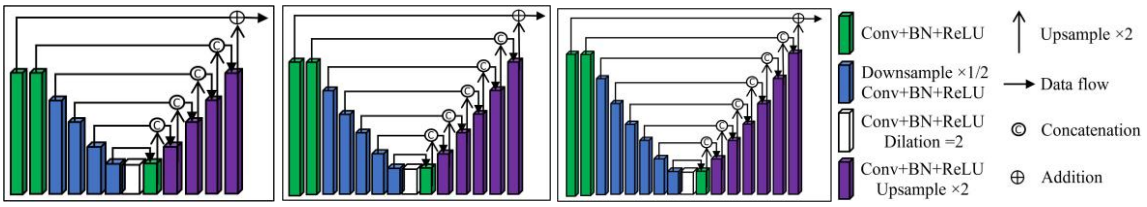


Fig. 1 The architecture of RSU block.



(a) RSU6 (b) RSU7 (c) RSU8

Fig. 2 The introduction of RSU6, RSU7 and RSU8.

extenuates the loss of fine details caused by direct upsampling with large scales.

3) A residual connection which fuses local features and the multi-scale features by the summation.

The residual U-blocks RSU6, RSU7 and RSU8 are utilized to embed into MODNet for going deeper. Take the RSU6 as an example (**Fig. 1**), the RSU6 consists of 6 encoder layers. A feature map is input to the first layer, which is a plain convolutional (Conv) layer following Bach Normal (BN) and ReLU, to extract the local feature. After that, 6 encoder layers extract features, each layer operates downsampling, Conv, BN and ReLU. The last layer (bottom green block) does not include downsampling but has dilated convolution with a dilation rate of 2. After the encoders, 6 decoder layers are used to recover features. Each decoder layer includes Conv, BN, ReLU and final upsampling. Each encoder layer is concatenated to corresponding similar scale decoder layers. The Internal channel number *M* is 32, which is similar to the MODNet internal channel in the high-resolution branch. The last encoder layer outputs feature map is added to the firstly of the RSU block. The output of the RSU6 block is a feature map which has the similar scale to the input feature map. The RSU7 and RSU8 have the similar structure to RSU6, but the encoder-decoder layers are 7 and 8 respectively. The three RSU blocks are shown in **Fig. 2**.

3.2 The architecture of RSUDISNet

The proposed theory builds the deeper high-resolution branch and fusion branch of MODNet by using the RSU block. The architecture is shown in **Fig. 3**. Suppose I is an input image. The low-resolution branch $S(I)$ architecture is MoblieNetV2 following the efficient atrous spatial pyramid polling (e-ASPP), which is an accurate semantic estimation. This branch is not built deeper. This branch predicts the coarse semantic mask s_p. It is supervised by a thumbnail of the ground truth matte α_g. Since s_p is supposed to be smooth, the L2 loss is used:

$$L_s = \frac{1}{2} \left\| s_p - G(\alpha_g) \right\|_2 \tag{1}$$

where G stands for 16× downsampling followed by Gaussian blur. It removes the fine structures (like hair) which are not essential to coarse semantic.

In the high-resolution branch D of the MODNet, the convolution layers are modified to become RSU7 and RSU6 layers. The outputs and inputs feature maps of each block have 64 channels. The RSU7 internal layer channel number is 32, and so is

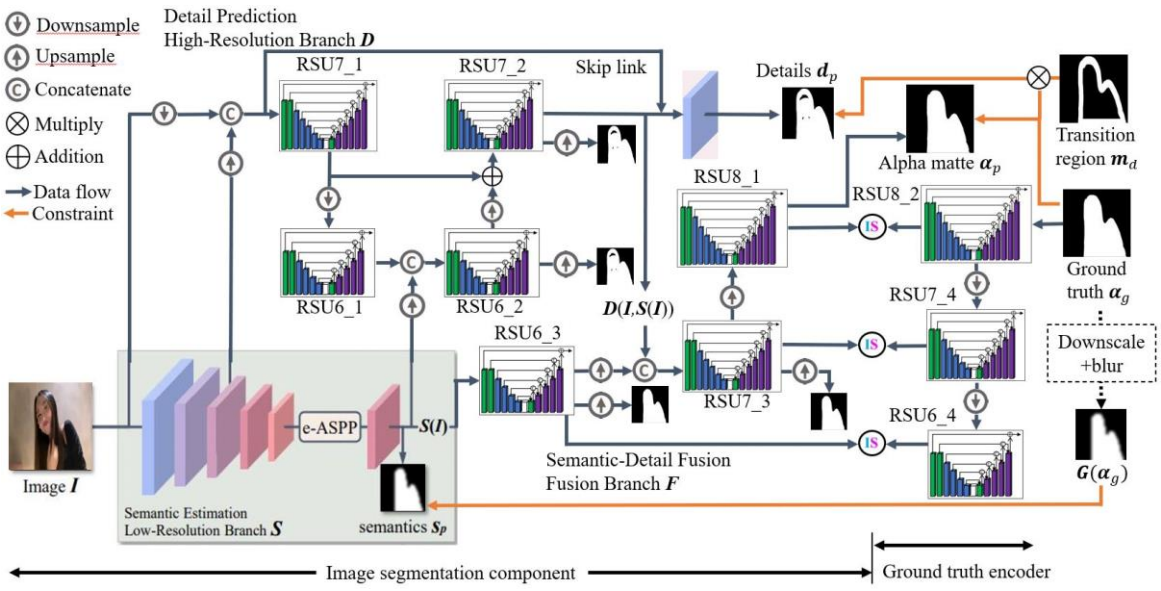


Fig. 3 The architecture of RSUDISNet

the RSU6. The resolution of the feature map is reduced to ¼ of the input image I in RSU6_1 and restored after the two blocks (RSU6_2 and RSU7_2). This makes the detail branch go deeper. However, the computing consumption is low and efficient.

The impact of the downsampling operation on D is negligible since it contains a skip link. The output of D is denoted as $D(I, S(I))$, which implies the dependency between sub-objectives – coarse semantic $S(I)$ is a priori for detail prediction. The boundary detail matte d_p from $D(I, S(I))$ and learn it through the L1 loss:

$$L_d^m = m_d \left\| d_p - \alpha_g \right\|_1 \quad (2)$$

where m_d is a binary mask to let L_d^m focus on the object boundaries. m_d is generated through dilation and erosion on α_g. Its values are 1 if the pixels are inside the transition region, and 0 otherwise. In addition, inspired by U2Net [8] loss function. The outputs of RSU6_2, and RSU7_2 are also rescaled to be similar resolution to m_d and are learned through L1 loss (L_d^x):

$$L_d^x = m_d \left\| d_p^x - \alpha_g \right\|_1 \quad (3)$$

The d_p^x denotes the predicted detail map, which is resized to be similar to the resolution of the ground truth α_g, from RSU blocks. The x represents the RSU block number – 6 and 7. The detail prediction loss is:

$$L_d = L_d^m + L_d^7 + L_d^6 \quad (4)$$

The fusion branch F combines semantics and details predictions. It includes RSU6, RSU7 and RSU8 blocks. The semantic prediction $S(I)$ is input to RSU6_3, the output feature is upsampled to match $D(I, S(I))$. The RSU6_3 output feature and $D(I, S(I))$ are concatenated. After that, the features input to RUS7_3, and then after upsampling is the RSU8_1. The output of RSU8_1 is transformed to one channel alpha matte α_p by a convolution layer following batch normal and ReLU. This is the final alpha matte which is constrained by:

$$L_\alpha^8 = \left\| \alpha_p^8 - \alpha_g \right\|_1 + L_c \quad (5)$$

where L_c is the composition loss [29]. It measures the absolute difference between input image I and the composited image obtained from α_p^8, the ground truth foreground, and the ground truth background.
In addition, the output features from RSU6_3 and RUS7_3 are also done in a 1x1 convolution and produce single-channel alpha mattes. They are rescaled to be similar resolution of the ground truth and be used to calculate the loss:

$$L_\alpha^x = \left\| \alpha_p^x - \alpha_g \right\|_1 + L_c \quad (6)$$

where α_p^x denotes the predicted alpha maps from RSU6_3 and RUS7_3. The final fusion-branch alpha loss is:

$$L_\alpha = L_\alpha^8 + L_\alpha^7 + L_\alpha^6 \quad (7)$$

3.3 The intermediate supervision (IS)

The intermediate supervision (IS) [9] is also used to train the fusion branch F. In the last section, the learning of the fusion branch F uses the single channel alpha matte map which is produced by convolving the last feature maps of particular deep layers from RSU8_1. This transforms the high-dimensional features to a single-channel alpha matte map which is a dimension reduction operation, inevitably losing critical cues. This leads the RSUDIS model easily to over-fit on the training. To avoid this issue, the intermediate supervision (IS) is utilized. Given an input image $I^{H\times W\times 3}$ and its corresponding alpha matte (ground truth) $\alpha_g^{W\times H\times 1}$, a self-supervised ground truth (GT) encoder is built and trained to extract the high-dimensional features using similar corresponding RSU blocks model F_{gt}. This is shown in **Fig. 3**, the ground truth encoder. All the RSU8_2, RSU7_4 and RSU6_4 are corresponding to RSU8_1, RSU7_3 and RSU6_3 respectively. The F_{gt} is learned by:

$$arg\min_{\theta_{gt}} \sum_{d=1}^{D} BCE(F_{gt}(\theta_{gt}, \alpha_g)_d, \alpha_g) \quad (8)$$

where θ_{gt} indicates the model weights, BCE is the binary cross entropy loss and D denotes the number of the intermediate feature maps. In this paper, D=3. After obtaining the GT encoder F_{gt}, its weights θ_{gt} are frozen for generating the "ground truth" high-dimensional intermediate deep features by:

$$f_D^G = F_{gt}^-(\theta_{gt}, G) \quad (9)$$

where $D = \{1, 2, 3\}$. The F_{gt}^- represents the F_{gt} without the last convolution layers for generating the single-channel alpha maps. F_{gt}^- is to supervise those corresponding features f_D^I from the RSUDISNet fusion branch F. In the fusion branch, each RSU block outputs high-dimensional intermediate feature maps f_D^I before producing the single-channel alpha maps. Each feature map f_d^I has the same dimension as its corresponding RSU block GT intermediate feature map f_d^G:

$$f_D^I = F_{sg}^-(\theta_{sg}, I) \quad (10)$$

where $D = \{1, 2, 3\}$. The θ_{sg} denotes the weights of the fusion branch. Then, the intermediate supervision (IS) via feature synchronization on the deep intermediate features can be conducted by the following high-dimensional feature consistency loss:

$$L_f = \sum_{d=1}^{D} \left\| f_d^I - f_d^G \right\|_2 \quad (11)$$

Finally, the loss of the RSUDISNet is

$$L = \lambda_s L_s + \lambda_d L_d + \lambda_\alpha L_\alpha + \lambda_f L_f \quad (12)$$

where λ_s, λ_d, λ_α and λ_f are hyper-parameters balancing the four losses. The train process is robust to these hyper-parameters. They are set to $\lambda_s = \lambda_\alpha = 3$, $\lambda_d = 10$ and $\lambda_f = 1$.

3.4 Data collection and labelling

The RSUDISNet architecture is introduced in the previous sections. The training of the RSUDISNet requires the RGB images and the corresponding alpha matte. The RGB images are extracted from the KOL videos. The 60 KOL videos are collected from different KOLs. Each video is extracted in 20 frames. They are from different times, having different poses and holding different objects.

The alpha matte kind mask is also important data for the training. Although this paper aim is to do KOL segmentation with holding objects, only providing a binary mask cannot train the high-resolution branch of the RSUDISNet. The alpha matte's transition region (value between 0 and 1) enables the high-resolution branch to be trained and work. However, the labelling of the alpha matte costs significant work. Therefore, a special alpha matte kind mask is created.

First, the RGB images are labelled with the shape of the KOL and the holding objects. Second, the labels are converted to binary masks. The KOL and holding objects are in the foreground, with the value of "1". The other thing on the RGB images is the background, with the value of "0". Finally, the binary mask extracts the boundary between foregrounds and backgrounds, then the boundary is set to the value of "0.5". In summary, the special training method with the special masks can train all three branches of the RUSDISNet.

4 Experimental results and discussion

In this section, The KOL dataset is used to compare RSUDISNet, and the original MODNet and U2Net. The dataset is split randomly to train, validation and test sets. The RSUDISNet, MODNet and U2Net are all trained with the KOL dataset. The performances of the three neural networks are tested by the test dataset. The segmentation results are compared with different thresholds from the alpha matte kind results. After the comparison, some virtual results are displayed.

4.1 Training of the neural networks

There are 1200 images from 60 KOL videos. They are randomly split into three sets: 1000 images for training, 100 for validation and 100 for test. The corresponding labelling is manually done by using Labelme. After labelling, the binary masks are created, and then the edge between foreground and background is added value of "0.5" with a thickness of 2 pixels. This is the ground truth $\boldsymbol{\alpha}_g$ for training. The

In the training of RSUDISNet, the transition region $\boldsymbol{m}_d$ is set to 30 pixels thickness. It is created from the edge of the ground truth. This is to enable the high-resolution branch of the RSUDISNet, which is the pre-process for the training. The images are resized to 512x512 for the input resolution. The training experiment shows that the learning rate 1e-4 is optimal. With this learning rate, the RSUDISNet is trained through 200 epochs. The 119th epoch indicates the best performance in the validation data.

The training of MODNet requires a similar pre-process for the training, because the MODNet is used for segmentation in there. The transition region for RUSDISNet is also used for MODNet training. The input resolution for MODNet training is set like this: the shorter side is set to 512 and the longer side is set following the image's original size ratio. For an example. If an image has a resolution of 1920x1080, the image is resized to 910x512. The optimal MODNet training learning rate is 1e-3. It also trained with 200 epochs. The 200th epoch has the best performance on the validation.

The U2Net architecture is different from the above two neural networks. It does not require the pre-processing for the training. The RGB images and the corresponding masks are used for U2Net

Table 1 The neural networks' size and test speed

Neural networks	Input resolution	Weights	FPS
U2Net	320×320	172 MB	35
MODNet	512×(width/height)	26 MB	54
RSUDISNet	512×512	30 MB	34

Table 2. Comparison of the three neural networks on KOL test data.

Models	U2Net			MODNet			RSUDISNet		
Threshold	mIoU	Acc	F_β	mIoU	Acc	F_β	mIoU	Acc	F_β
0.2	0.8275	0.9208	0.8632	0.8679	0.9693	0.9087	0.8759	0.9710	0.9075
0.3	0.8297	0.9216	0.8697	0.8710	0.9701	0.9123	0.8812	0.9725	0.9132
0.4	**0.8322**	0.9222	0.8754	0.8733	0.9707	0.9152	0.8852	0.9736	0.9178
0.5	0.8319	**0.9228**	**0.8778**	0.8751	0.9711	0.9178	0.8884	0.9744	0.9218
0.6	0.8313	0.9226	0.8767	**0.8759**	**0.9713**	0.9200	0.8911	0.9752	0.9257
0.7	0.8309	0.9227	0.8756	0.8756	0.9712	**0.9216**	0.8933	0.9758	0.9297
0.8	0.8306	0.9224	0.8740	0.8743	0.9710	0.9206	**0.8943**	**0.9761**	0.9339
0.9	0.8284	0.9219	0.8719	0.8719	0.9704	0.9187	0.8922	0.9758	**0.9382**

training. The input resolution of U2Net is 320x320. The optimal learning rate is 0.01. The U2Net is trained through 150 epochs. The 148th epoch performs the best on the validation dataset.

To compare the three neural network performances, they applied to the test dataset with different thresholds from 0.1 to 0.9. The comparison is shown in the next section.

4.2 Results and discussions

The processing time of U2Net, MODNet and RSUDISNet is shown in **Table 1**. The RSUDISNet weights volume is about 30 MB which does not increase significantly from MODNet (26 MB). The U2Net which is used for testing in this paper is the heavy weights one (172 MB), which is claimed the best performance in [8]. The GPU system used for the test is NVIDIA 1070Ti, 8GB. The RSUDISNet is slower 1/3 than MODNet, but similar to U2Net. The processing time is near real time, if the KOL video is 30 frame rates. The RSUDISNet has a deeper architecture than MODNet but does not increase processing time seriously.

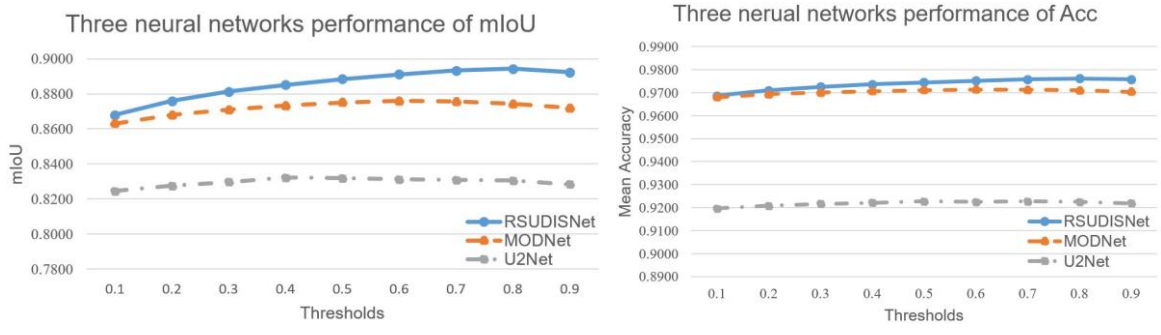
The proposed RSUDISNet performance is compared with MODNet and U2Net. The 100 test images are used to do the evaluation. The evaluation standard matrix is pixel mask accuracy (Acc), mean of intersection over union (mIoU) and F_β:

$$F_\beta = \frac{(1+\beta^2)\text{Precision}\times\text{Recall}}{\beta^2\text{Precision}+\text{Recall}} \tag{13}$$

where β=0.3. F_β is the score for the single threshold on a whole test dataset. The outputs of the three neural networks have the alpha matte kind masks. For segmentation evaluation, the threshold from 0.2 to 0.9 is tested.

Table 2 shows the evaluation results. For all three neural networks, the threshold's changing does not affect the performance significantly. In U2Net, the best performance of mIoU is shown in threshold >0.4, and the best Acc and F_β is at threshold >0.5. In MODNet, the threshold >0.6 shows the best mIoU and Acc. The threshold >0.7 shows the best F_β. The RSUDISNet has the best mIoU and Acc in threshold >0.8. The best F_β is shown in threshold >0.9. The RSUDISNet has higher values than U2Net and MODNet. Even at a threshold >0.5, the evaluation values are slightly better than MODNet. The U2Net performs the worst results. **Fig. 4** displays the performance of the three neural networks on mIoU and Acc. The RSUDISNet outperforms the MODNet and U2Net.

The RSUDISNet architecture includes a low-resolution branch which is similar to the low-resolution branch of MODNet. However, the high-resolution branch and the fusion branch are deeper than the MODNet. Therefore, the RSUDISNet performs better than MODNet. The deeper architecture is inspired by the RSU blocks from U2Net. In the high-resolution branch of RSUDISNet, two levels of the RSU blocks, RSU6 and RSU7 are utilized, which extract features from the detail part



(a) mIoU (b) Acc

Fig. 4 Comparison of three neural networks on mIoU and Acc with KOL test data.

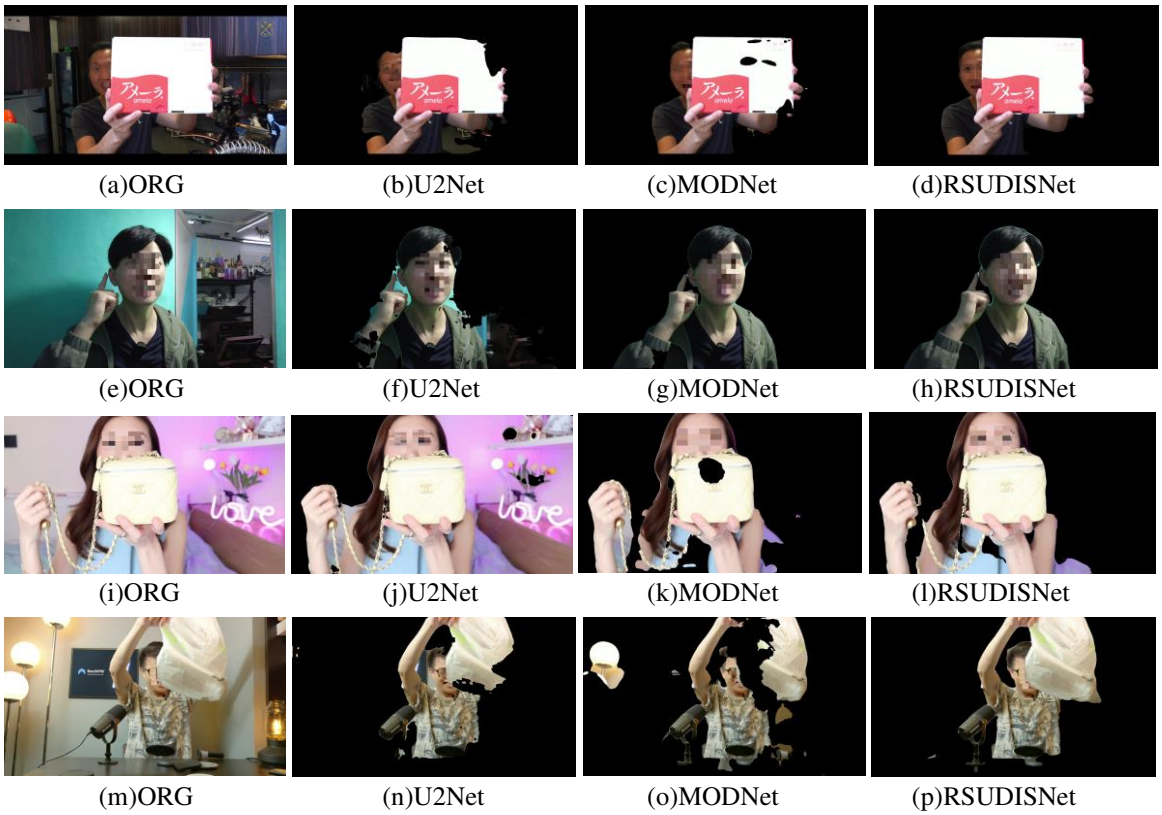


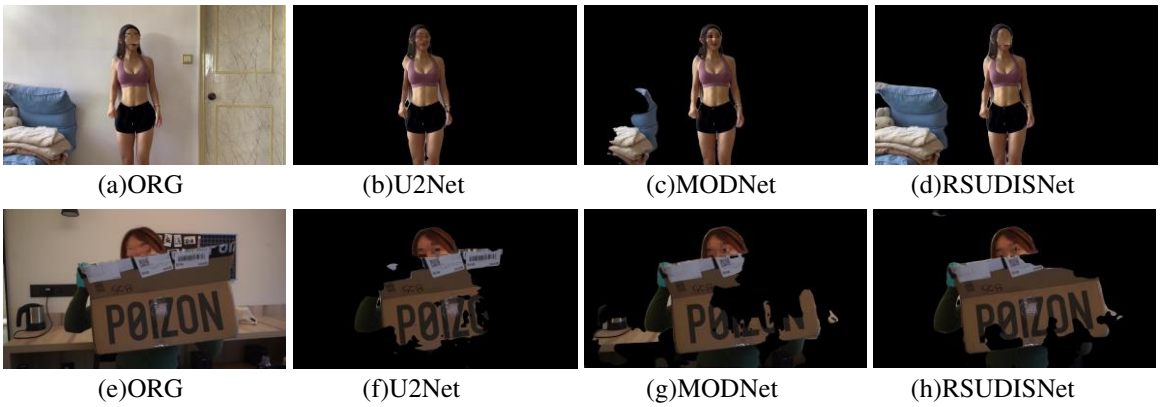
Fig. 5. The KOL image results of three neural networks. From the left, first column is original images (ORG), second column is results of U2Net, third column is MODNet's results, fourth column is RSUDISNet's results.

of the image with different scales. In the fusion branch, three levels of RSU blocks are used: RSU6, RSU7 and the supposed RSU8. These three levels of blocks are only up-scale feature maps to the input resolution of the original input image. Finally, the IS training method reduces the overfitting affection. The U2Net's deeper architecture makes it easy to get overfitting [9], because the deeper the features, the more complicated the model. Therefore, on the 1100 images training, the performance is worse than RSUDISNet.

Fig. 5 displays KOL segmentation virtual results by using the three neural networks. The actual outputs of the three neural networks are binary masks after the threshold >0.5. In the figure, in order to highlight the segmentation results more obviously, the neural network results display the foreground detection with the original colour. The predicted background is shown in black colour. From the left column to the right column, they are the original RGB images, U2Net predicts, MODNet predicts and RSUDISNet predicts.

It has been demonstrated that the U2Net responds well to saliency object detection. However, the KOL and the holing products are rarely shown as salient items in the supposed KOL dataset images. **Fig. 5** shows that there is a KOL segmentation issue with the U2Net. The human hair in **Fig. 5**b is poorly segmented because the background colour is similar to the hair colour. As the background patterns and colours **Fig. 5** f and j differ, the U2Net segmentation has a false positives region. In particular, in **Fig. 5**j, the KOL is segmented as a single foreground object within the purple background region. **Fig. 5**n yields a better result than the three images above; the plastic bag is nearly segmented, despite a few false negatives. However, there is a poor division of segmentation between the man's arm and head. Another possible reason for the U2Net issue is the small size of the dataset. The U2Net architecture includes nested U-structure. It may be too deep to handle the overfitting with a small dataset and this requires future testing.

The MODNet is proved to segment well to the human portrait. However, the KOL dataset in this paper includes many humans and their holding products. **Fig. 5**g shows that the MODNet segments the KOL well, if the image includes only humans, although there is a small region of false negative under the KOL's arm. In **Fig. 5**c, k and o, the KOLs are holding different products which affect the MODNet segmentation results. There are small false positive regions on **Fig. 5**c and k. Particularly, the plastic bag in **Fig. 5**o is segmented worse than U2Net.



(a)ORG (b)U2Net (c)MODNet (d)RSUDISNet

(e)ORG (f)U2Net (g)MODNet (h)RSUDISNet

Fig. 6. The more segmentation results of three neural networks. From the left, first column is original images (ORG), second column is results of U2Net, third column is MODNet's results, fourth column is RSUDISNet's results.

Compared to the U2Net and MODNet, the RSUDISNet's segmentation results on **Fig. 5**d and h are close to perfect. The Human is segmented better than MODNet. The object is segmented well. The RSUDISNet is deeper than MODNet, so it performs better. In **Fig. 5**l, the RSUDISNet segments the Shoulder bag better than MODNet, because the RSU block has a good response on saliency objects. The KOL lady is still segmented with some false positives and false negatives; however, the background is removed better than U2Net. This may be because the decomposition loss has good responses on human portrait shape, which overcomes different colours of background. **Fig. 5**p shows the plastic bag is segmented better than the other two neural networks. In the test dataset, most of the KOL images show RSUDISNet segmenting better results. This can also prove that the IS overcomes the overfitting significantly.

However, there are still some images that show the limitation of the RSUDISNet segmentation. **Fig. 6** displays some results that are not good. The first raw of **Fig. 6**b, c and d indicate that the RSUDISNet segments more about the sofa which is not supposed to be segmented. In these figures, the U2Net performs the best. The KOL is segmented as foreground. The MODNet segments part of the sofa, and the false positive region is smaller than RSUDISNet. The possible reason for this RSUDISNet problem is that the RSUDISNet has some MODNet structures, such as low-resolution branch and decomposition loss. However, this requires more tests to analyse the actual reason.

In the second raw of Fig. 6f, the U2Net segments the KOL's face and arm, but the paper box with big region of false negatives. The MODNet segments the KOL and paper box with more regions (**Fig. 6**g). Nevertheless, it gets some false positives which are the regions of the table. The result of RSUDISNet in **Fig. 6**h is better than the results of U2Net and MODNet. However, the segmentation is still not perfect. The RSUDISNet is built by the combination of U2Net and MODNet. Therefore, the limitation of the RSUDISNet is from the disability of the two neural networks, although the IS has the contribution of overfitting overcoming.

5 Conclusion

In this paper, the proposed RSUDISNet combing U2Net and MODNet works effectively. RSUDISNet also incorporates IS to overcome overfitting. The RSUDISNet makes the high-resolution branch and fusion branch much deeper than MODNet by using RSU blocks from U2Net. The weights are 30MB in size, which is slightly bigger than MODNet's 26MB. Although the running speed is slower than MODNet, it is still close to real-time on KOL social videos. The experiemntal results show that the RSUDISNet outperforms U2Net and MODNet. It gives good performance on the application of KOL segmentation on social media videos.

There are some limitations in the RSUDISNet. One of them is due to the problems of MODNet affecting the RSUDISNet performance. For example, **Fig. 6**d shows the result of RSUDISNet is worse than U2Net and MODNet. In this case, the U2Net result is good, but the MODNet has a false positive problem. This requires more analysis to understand the reason. Secondly, the original problems from both two neural networks, MODNet and U2Net, may lead RSUDISNet to perform not well. For instance, although **Fig. 6**h indicates the RSUDISNet result is better than U2Net and MODNet, the segmentation still has some false negatives. This is because both MODNet and U2Net cannot segment the paper box.

6 Statements and Declarations

Availability of data and materials The data and materials during the current study are available from the corresponding author on reasonable request.

Author contributions All authors contributed to the study conception and design. The methodology and experiment were performed by Cheng Yang. Dataset creation was completed by Fucheng Zheng. The conceptualization and supervising the progress was done by Patrick Lam and Peter Han Joo Chong. The first draft of the manuscript was written by Cheng Yang and Duaa Zuhair Al-Hamid. Peter Han Joo Chong revised and commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Competing Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E.: Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience,* 2018, 7068349 (2018). https://doi.org/10.1155/2018/7068349
2. Gruosso, M., Capece, N., Erra, U.: Human segmentation in surveillance video with deep learning. Multimedia Tools and Appl. 80, 1175-1199 (2021)
3. Lin, K., Wang, L., Luo, K., Chen, Y., Liu, Z., Sun M.T.: Cross-domain complementary learning using pose for multi-person part segmentation. IEEE Transactions on Circuits and Systems for Video Technology. 31(3), 1066-1078 (2020)
4. Sun, Y., Tang, C.K., Tai, Y.W.: Human instance matting via mutual guidance and multi-instance refinement. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2647-2656 (2022)
5. Ke Z., Sun J., Li K., Yan Q., Lau R. W.: Modnet: Real-time trimap-free portrait matting via objective decomposition. The AAAI Conference on Artificial Intelligence. 36(1), 1140-1147 (2022)
6. Liu J.J., Hou Q., Liu Z.A., Cheng M.M.: Poolnet+: Exploring the potential of pooling for salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 45(1), 887-904 (2022)
7. Ke Y. Y., Tsubono T.: Recursive contour-saliency blending network for accurate salient object detection. The IEEE/CVF Winter Conference on Applications of Computer Vision, 2940-2950 (2022)
8. Qin X., Zhang Z., Huang C., Dehghan M., Zaiane O. R., Jagersand M.: U2-Net: Going deeper with nested U-structure for salient object detection. Pattern recognition. 106, 107404 (2020)
9. Qin X., Dai H., Hu X., Fan D.P., Shao L., Van Gool L.: Highly accurate dichotomous image segmentation. Computer Vision–ECCV 2022: 17th European Conference. Springer, Tel Aviv, Israel. 38-56 (2022)
10. Koonce B.: VGG Network. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization. 35-50 (2021)
11. Luo Y., Zheng Z., Zheng L., Guan T., Yu J., Yang Y.: Macro-micro adversarial network for human parsing. The European conference on computer vision (ECCV). 418-434 (2018)
12. Yang L., Song Q., Wang Z., Jiang M.: Parsing r-cnn for instance-level human analysis. The IEEE/CVF Conference on Computer Vision and Pattern Recognition. 364-373 (2019)
13. Fang H.S., Lu G., Fang X., Xie J., Tai Y.W., Lu C.: Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. arXiv preprint arXiv:1805.04310. https://arxiv.org/abs/1805.04310 (2018). Accessed 11 May 2018
14. Shen X., Tao X., Gao H., Zhou C., Jia J.: Deep automatic portrait matting. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands Proceedings, Part I 14, Springer. 92-107 (2016)
15. Lepcha, D. C., Goyal, B., Dogra, A.: Image matting: a comprehensive survey on techniques, comparative analysis, applications and future scope. International Journal of Image and Graphics. 23(01), 2350011 (2023). https://doi.org/10.1142/S0219467823500110.
16. Sun Y., Tang C.K., Tai Y.W.: Semantic image matting. The IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11120-11129 (2021)
17. Wu X., Fang X.-N., Chen T., Zhang F.L.: JMNet: A joint matting network for automatic human matting: Computational Visual Media. 6, 215-224 (2020)
18. Liu J., Yao, Y., Hou, W., Cui, M., Xie, X., Zhang, C., Hua, X.S.: Boosting semantic human matting with coarse annotations. The IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8563-8572 (2020)
19. Chen, X., Zhu, Y., Li, Y., Fu, B., Sun, L., Shan, Y., Liu, S.: Robust Human Matting via Semantic Guidance. The Asian Conference on Computer Vision. 2984-2999 (2022)
20. Lin, S., Yang, L., Saleemi I., Sengupta, S.: Robust high-resolution video matting with

temporal guidance. The IEEE/CVF Winter Conference on Applications of Computer Vision. 238-247 (2022)
21. Cong, R., Yang, N., Li, C., Fu, H., Zhao, Y., Huang, Q. Kwong, S.: Global-and-local collaborative learning for co-salient object detection. IEEE transactions on cybernetics. (2022)
22. Li, G., Liu, Z., Zeng, D., Lin, W., Ling, H.: Adjacent context coordination network for salient object detection in optical remote sensing images. IEEE Transactions on Cybernetics. 53(1), 526-538 (2022)
23. Xu, B., Liang, H., Liang, R., Chen P.: Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. The AAAI Conference on Artificial Intelligence. 35(4), 3004-3012 (2021)
24. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand M.: Basnet: Boundary-aware salient object detection. The IEEE/CVF conference on computer vision and pattern recognition. 7479-7489 (2019)
25. Lee, M. S., Shin W., Han, S. W.: TRACER: Extreme Attention Guided Salient Object Tracing Network (Student Abstract). The AAAI Conference on Artificial Intelligence. 36(11), 12993-12994 (2022)
26. Cheng, M.M., Gao, S.H., Borji, A., Tan, Y.Q., Lin, Z., Wang, M.: A highly efficient model to study the semantics of salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 44(11), 8006-8021 (2021)
27. GongyangLi, Z., Bai, Z., Lin, W., Ling, H.: Lightweight Salient Object Detection in Optical Remote Sensing Images via Feature Correlation. IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. 60, 5617712 (2022)
28. Wang, Z., Zhang, Y., Liu, Y., Zhu, D., Coleman, S. A., Kerr, D.: ELWNet: An Extremely Lightweight Approach for Real-Time Salient Object Detection. IEEE Transactions on Circuits and Systems for Video Technology. (2023). https://ieeexplore.ieee.org/document/10107621
29. Xu, N., Price, B., Cohen, S., Huang T.: Deep image matting. The IEEE conference on computer vision and pattern recognition. 2970-2979 (2017)